

Online Expectation Maximization based algorithms for inference in Hidden Markov Models

Sylvain Le Corff^{*†} and Gersende Fort[†]

February 27, 2012

Abstract

The Expectation Maximization (EM) algorithm is a versatile tool for model parameter estimation in latent data models. When processing large data sets or data stream however, EM becomes intractable since it requires the whole data set to be available at each iteration of the algorithm. In this contribution, a new generic online EM algorithm for model parameter inference in general Hidden Markov Model is proposed. This new algorithm updates the parameter estimate after a block of observations is processed (online). The convergence of this new algorithm is established, and the rate of convergence is studied showing the impact of the block size. An averaging procedure is also proposed to improve the rate of convergence. Finally, practical illustrations are presented to highlight the performance of these algorithms in comparison to other online maximum likelihood procedures.

1 Introduction

The Expectation Maximization (EM) algorithm is a well-known iterative algorithm to solve maximum likelihood estimation in incomplete data models [12]. In this context, model parameter estimates are obtained by maximizing the log-likelihood of the observations $Y_{0:T}$. Despite in incomplete data models the log-likelihood is not explicit, EM algorithm is generally simple to implement since it relies on complete data computations: each iteration consists in a E-step where the expectation of the complete log-likelihood under the conditional distribution of the latent data given the observations is computed; and a M-step, which updates the parameter estimate based on this conditional expectation.

In many situations of interest, the complete data likelihood belongs to the exponential family. In this case, the E-step consists in the computation of

^{*}This work is partially supported by the French National Research Agency, under the program ANR-08-BLAN-0218 BigMC

[†]LTCI, CNRS and TELECOM ParisTech, 46 rue Barrault 75634 Paris Cedex 13, France

the expectation of the complete data sufficient statistic under the conditional distribution. In such case, the EM algorithm can be considered equivalently as an iterative algorithm in the space of the complete data sufficient statistics (instead of in the parameter space).

The EM algorithm has been successfully applied for maximum likelihood inference in general state-space models. Except for simple models the E-step is intractable and has to be approximated e.g. by Monte Carlo methods such as Markov Chain Monte Carlo methods or Sequential Monte Carlo methods (see resp. [5, 17]) depending on the complexity of the model.

When processing large data sets or data streams however, the EM algorithm might become impractical. *Online* variants of the EM algorithm have been first proposed for independent and identically distributed (i.i.d.) observations. The first online procedure for parameter estimation was introduced in [29] by Titterton. This algorithm relies on a stochastic gradient approach which aims at incorporating the newly available observation. In Cappé and Moulines [4], the proposed algorithm is more closely related to the original EM recursion: in the case of an exponential complete-data likelihood, the E-step is replaced by a stochastic approximation step while the M-step remains unchanged.

More complex incomplete data models such as Hidden Markov Models (HMM) are of common use to represent time series in many fields such as statistics, information engineering and financial econometrics, see [15, 31]. An online version of the EM algorithm for inference in HMM when both the observations and the states take a finite number of values (resp. when the states take a finite number of values) was recently proposed by Mongillo and Denève [24] (resp. Cappé [3]). In Cappé [3], the algorithm relies on the ability to compute approximations of the filtering distribution and on an intermediate quantity based on the sufficient statistics. In order to update these computations recursively, stochastic approximation procedures are introduced. This algorithm has been extended to the case of general state-space models by substituting deterministic approximation of the smoothing probabilities for Sequential Monte Carlo algorithms (see Cappé [2], Del Moral *et al.* [8] and Le Corff *et al.* [22]).

Despite the encouraging first results when applying these online EM algorithms, the convergence of these algorithms and the characterization of the limit points (when the number of observations tends to infinity) remain an open question. The convergence of the online variants of the EM algorithm for i.i.d. observations is addressed by Cappé and Moulines [4]: the limit points are the stationary points of the Kullback-Leibler divergence between the marginal distribution of the observation and the model distribution. There do not exist convergence results for the online EM algorithms for general state-space models (some insights on the asymptotic behavior are nevertheless given in Cappé [3]): the introduction of many approximations at different steps of the algorithms makes the analysis quite challenging.

In this contribution, a new online EM algorithm is proposed for HMM with exponential complete-data likelihood. It sticks more closely to the principles of the original batch-mode EM algorithm. The M-step (and thus, the update of the parameter) occurs at some deterministic times $\{T_k\}_{k \geq 1}$ i.e. we propose

to keep a fixed parameter estimate for blocks of observations of increasing size. More precisely, let $\{T_k\}_{k \geq 0}$ be an increasing sequence of integers ($T_0 = 0$). For each $k \geq 0$, the parameter's value is kept fixed while accumulating the information brought by the observations $Y_{T_k+1:T_{k+1}}$. Then, the parameter is updated at the end of the block. This algorithm is an online algorithm since the sufficient statistics of the k -th block can be computed on the fly by updating an intermediate quantity when a new observation Y_t , $t \in \{T_k + 1, \dots, T_{k+1}\}$ is available. Such recursions are provided in recent works on online estimation in HMM, see [2, 3, 8].

This new algorithm, called *Block Online EM* algorithm (BOEM) is derived in Section 2 together with an *averaged* version. Section 3 is devoted to practical applications: BOEM is used to perform parameter inference in HMM where the forward recursions mentioned above are available explicitly (this occurs e.g. for finite state-space HMM). In the case of finite state-space HMM, BOEM is compared to a gradient-type recursive maximum likelihood procedure and to the online EM of [3].

The convergence of BOEM is addressed in Section 4. BOEM is seen as a perturbation of a deterministic *limiting EM* algorithm, the limiting behavior of which is studied through a Lyapunov-function technique. The perturbation is shown to vanish (in some sense) as the number of observations increases thus implying that BOEM inherits the asymptotic behavior of the *limiting EM* algorithm. Finally, in Section 5, we prove that the rate of convergence of BOEM strongly depends upon the block size sequence: this rate is optimal when the block size increases exponentially which is, quite unfortunately, of poor practical interest. Nevertheless, we prove that the averaged BOEM reaches this optimal rate of convergence for slowly increasing block size sequence. All the proofs are postponed in Section 6; supplementary materials are provided in [21].

2 The Block Online EM algorithms

2.1 Notations and Model assumptions

Let $\mathbf{Y} = \{Y_t\}_{t \in \mathbb{Z}}$ be the observation process defined on $(\Omega, \mathbb{P}_*, \mathcal{F})$ and taking values in $\mathbb{Y}^{\mathbb{Z}}$ where \mathbb{Y} is a general space endowed with a countably generated σ -field $\mathcal{B}(\mathbb{Y})$.

A HMM model parameterized by θ , for θ in a set $\Theta \subseteq \mathbb{R}^{d_\theta}$, is fitted to the observations: consider a family of transition kernels $\{m_\theta(x, x') d\lambda(x')\}_{\theta \in \Theta}$ onto $\mathbb{X} \times \mathcal{B}(\mathbb{X})$ where \mathbb{X} is a general state-space equipped with a countably generated σ -field $\mathcal{B}(\mathbb{X})$, and λ is a bounded non-negative measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. Let $\{g_\theta(x, y) d\nu(y)\}_{\theta \in \Theta}$ be a family of transition kernels on $(\mathbb{X} \times \mathcal{B}(\mathbb{Y}))$, where ν is a measure on $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$.

For any initial distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, any $\theta \in \Theta$, any $r < s \leq t$ and

any sequence $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, define the probability measure $\Phi_{\theta,s,t}^{\chi,r}(\cdot, \mathbf{y})$ by

$$\Phi_{\theta,s,t}^{\chi,r}(h, \mathbf{y}) \stackrel{\text{def}}{=} \frac{\int \chi(dx_r) \left\{ \prod_{i=r}^{t-1} m_{\theta}(x_i, x_{i+1}) g_{\theta}(x_{i+1}, y_{i+1}) \right\} h(x_{s-1}, x_s, y_s) d\lambda(x_{r+1:t})}{\int \chi(dx_r) \left\{ \prod_{i=r}^{t-1} m_{\theta}(x_i, x_{i+1}) g_{\theta}(x_{i+1}, y_{i+1}) \right\} d\lambda(x_{r+1:t})}, \quad (1)$$

for any bounded function h and where, for any $r \leq t$, we will use the shorthand notation $x_{r:t}$ for the sequence (x_r, \dots, x_t) . Note that if $\{(X_t, Y_t)\}_{t \in \mathbb{Z}}$ is a HMM with transition kernels m_{θ} and g_{θ} , $\Phi_{\theta,s,t}^{\chi,r}(h, Y)$ is the conditional expectation of $h(X_{s-1}, X_s, Y_s)$ given $Y_{r+1:t}$ when $X_r \sim \chi$:

$$\Phi_{\theta,s,t}^{\chi,r}(h, \mathbf{Y}) = \mathbb{E}_{\theta} [h(X_{s-1}, X_s, Y_s) | Y_{r+1:t}], \quad X_r \sim \chi. \quad (2)$$

It is assumed that the HMM is *exponential* i.e.

- A1** (a) There exist continuous functions $\phi : \Theta \rightarrow \mathbb{R}$, $\psi : \Theta \rightarrow \mathbb{R}^d$ and $S : \mathbb{X} \times \mathbb{X} \times \mathbb{Y} \rightarrow \mathbb{R}^d$ s.t.

$$\log m_{\theta}(x, x') + \log g_{\theta}(x', y) = \phi(\theta) + \langle S(x, x', y), \psi(\theta) \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product on \mathbb{R}^d .

- (b) There exists an open subset \mathcal{S} of \mathbb{R}^d that contains the convex hull of $S(\mathbb{X} \times \mathbb{X} \times \mathbb{Y})$.
- (c) There exists a continuous function $\bar{\theta} : \mathcal{S} \rightarrow \Theta$ s.t. for any $s \in \mathcal{S}$,

$$\bar{\theta}(s) = \operatorname{argmax}_{\theta \in \Theta} \{ \phi(\theta) + \langle s, \psi(\theta) \rangle \}.$$

2.2 Block Online EM (BOEM)

Define

$$\bar{S}_{\tau}^{\chi,T}(\theta, \mathbf{Y}) \stackrel{\text{def}}{=} \frac{1}{\tau} \sum_{t=T+1}^{T+\tau} \Phi_{\theta,t,T+\tau}^{\chi,T}(S, \mathbf{Y}). \quad (3)$$

Once again, note that if $\{(X_t, Y_t)\}_{t \in \mathbb{Z}}$ is a HMM with transition kernels m_{θ} and g_{θ} , $\bar{S}_{\tau}^{\chi,T}(\theta, \mathbf{Y})$ is the conditional expectation of the additive functional $\sum_{t=T+1}^{T+\tau} S(X_{t-1}, X_t, Y_t)$ given $Y_{T+1:T+\tau}$ when $X_T \sim \chi$:

$$\bar{S}_{\tau}^{\chi,T}(\theta, \mathbf{Y}) = \frac{1}{\tau} \sum_{t=T+1}^{T+\tau} \mathbb{E}_{\theta} [S(X_{t-1}, X_t, Y_t) | Y_{T+1:T+\tau}], \quad X_T \sim \chi.$$

Note that (3) can be computed without any storage of the observations on the block: the algorithm is not faced to any memory capacity issue, this is a streaming procedure (see Section 2.4 below).

Let $\{\tau_n\}_{n \geq 1}$ be a sequence of positive integers and set

$$T_n \stackrel{\text{def}}{=} \sum_{k=1}^n \tau_k \quad \text{and} \quad T_0 \stackrel{\text{def}}{=} 0; \quad (4)$$

τ_n denotes the length of the block n . To ensure the stability of this stochastic iterative algorithm, we use a reprojection scheme adapted from [6]. Let $\{\Theta_n\}_{n \geq 0}$ be a sequence of compact subsets of Θ s.t.

$$\forall n \geq 0, \Theta_n \subset \Theta_{n+1} \quad \text{and} \quad \Theta = \bigcup_{n \geq 0} \Theta_n. \quad (5)$$

Given an initial value $\theta_0 \in \Theta_0$ and starting with $p_0 = 0$, the BOEM algorithm defines a sequence $\{\theta_n\}_{n \geq 1}$ by

$$\begin{aligned} \theta_{n-1/2} &\stackrel{\text{def}}{=} \bar{\theta} [\bar{S}_{\tau_n}^{\chi, T_{n-1}}(\theta_{n-1}, \mathbf{Y})], \\ \theta_n &= \begin{cases} \theta_{n-1/2} & \text{if } \theta_{n-1/2} \in \Theta_{p_n} \\ \theta_0 & \text{otherwise and set } p_n = p_{n-1} + 1. \end{cases} \end{aligned} \quad (6)$$

p_n counts the number of truncations; it is proved in Theorem 4.4 that $\{p_n\}_{n \geq 0}$ is finite w.p.1. i.e. w.p.1., $\theta_n = \theta_{n-1/2}$ for all n large enough. BOEM updates the parameter estimates by using the integrals (3) computed on non-overlapping block of observations; the expectation is with respect to (w.r.t.) a conditional distribution given the (random) observations $Y_{T+1:T+\tau}$. Consequently, it is a stochastic iterative algorithm.

For ease of notation, it is assumed in this recursion that the initial distribution χ is the same for all blocks even though it will be clear in Section 4 that the initial distribution can change over blocks. We will choose a positive sequence $\{\tau_n\}_{n \geq 1}$ s.t. $\lim_{n \rightarrow \infty} \tau_n = +\infty$. Indeed, we will prove that $\lim_{\tau \rightarrow \infty} \bar{S}_{\tau}^{\chi, T}(\theta, \mathbf{Y})$ exists \mathbb{P}_\star - a.s (see Theorem 4.1 below), and it is thus expected that BOEM applied with such a sequence $\{\tau_n\}_{n \geq 1}$ will have the same asymptotic behavior as the iterative procedure in which $\bar{S}_{\tau_n}^{\chi, T_{n-1}}(\theta_{n-1}, \mathbf{Y})$ is replaced by its limit. We will give a rigorous proof of this intuition in section 4, as well as rigorous assumptions on $\{\tau_n\}_{n \geq 1}$.

2.3 Averaged Block Online EM

When τ_n is large, $\bar{S}_{\tau_n}^{\chi, T}(\theta, \mathbf{Y})$ may be seen as an estimator of the a.s. limit $\lim_{\tau \rightarrow \infty} \bar{S}_{\tau}^{\chi, T}(\theta, \mathbf{Y})$. By analogy to the regression problem, an estimator with reduced variance can be obtained by averaging and weighting the successive estimates (see [19, 26] for a discussion on the averaging procedures). Define $\Sigma_0 \stackrel{\text{def}}{=} 0$ and for $n \geq 1$,

$$\Sigma_n \stackrel{\text{def}}{=} \frac{1}{T_n} \sum_{j=1}^n \tau_j \bar{S}_{\tau_j}^{\chi, T_{j-1}}(\theta_{j-1}, \mathbf{Y}); \quad (7)$$

note that this quantity can be computed iteratively and does not require to store the past statistics $\bar{S}_{\tau_j}^{\chi, T_{j-1}}$. Given an initial value $\tilde{\theta}_0$, the averaged BOEM algorithm defines a sequence $\{\tilde{\theta}_n\}_{n \geq 1}$ by

$$\tilde{\theta}_n \stackrel{\text{def}}{=} \bar{\theta}(\Sigma_n). \quad (8)$$

2.4 Comments on the implementation of the algorithms

About the initial distribution χ : in (3), the computation on each block is performed with the same initial distribution χ . This simplifies the presentation of the algorithms and reinforces the readability of the proofs. Time dependent initial distributions could be considered, such as using the filtering distribution obtained at the end of block n to initialize block $n + 1$. In this case, the initial distribution depends on the past observations and the current parameter. Different strategies are numerically compared in Section 3.

The asymptotic behavior of our algorithms is derived under so-called *strong mixing conditions* of the hidden chain: this implies the forgetting of the initial condition at a geometric rate, uniform in the initial distribution χ . We prove asymptotic results for the algorithms described above (i.e. with a fixed initial distribution χ) but our results remain true for time dependent initialization strategies. Details are omitted.

Streaming: our algorithms update the parameter after processing a block of observations. Nevertheless, the intermediate quantity $\bar{S}_{\tau_n}^{\chi, T_{n-1}}(\theta_{n-1}, \mathbf{Y})$ can be either exactly computed or approximated in such a way that the observations are processed online. In this case, once received, an observation is used to update the intermediate quantity and then removed from the memory. The exact computation is detailed in [3, Section 2.2] and [8, Proposition 2.1] and can be applied e.g. to finite state-space HMM. [8] proposed a Sequential Monte Carlo approximation to perform this update online for more complex models (the convergence of BOEM combined with this method is addressed in [20]). Therefore, BOEM, its averaged version and its particle approximations can be described as streaming algorithms.

About the block size $\{\tau_n\}_{n \geq 1}$: it is expected that, when the number of observations tends to infinity, BOEM behaves like a *limiting EM*, i.e. an EM procedure with an infinity of observations. In Section 4, we characterize the asymptotic behavior of this *limiting EM* algorithm and, in order to inherit this asymptotic behavior, the number of observations per block used in BOEM has to increase to infinity. We will see in Sections 4 and 5 that polynomially increasing sizes $\tau_n \sim cn^a$ ($a > 1$) are enough. On a practical point of view, τ_n can be constant for the first iterations so that the parameter is updated sufficiently enough in the first part of the run. Then, τ_n increases like cn^a and the user can choose c in such a way that the block sizes do not grow too fast. The influence of the block sizes on the convergence of BOEM and its averaged version are illustrated in Section 3 (see also Section 5 for the computation of the rates of convergence).

3 Application to inverse problems in Hidden Markov Models

In Section 3.1, the performance of BOEM and its averaged version are illustrated in a linear Gaussian model. In this case, the quantity $\bar{S}_{\tau_n}^{\chi, T_{n-1}}(\theta_{n-1}, \mathbf{Y})$ can be exactly computed using a Kalman smoother but this requires to store the data on each block. $\bar{S}_{\tau_n}^{\chi, T_{n-1}}(\theta_{n-1}, \mathbf{Y})$ can also be approximated using a streaming procedure (i.e. without storing any data, see [20] for Sequential Monte Carlo methods) without modifying the limiting behavior of the algorithm. In the experiment below, we use the Kalman smoother. The role of the block size $\{\tau_n\}_{n \geq 1}$ and of the initialization scheme are discussed in Section 3.1.

In Section 3.2, BOEM is compared to online maximum likelihood procedures in the case of finite state-space HMM.

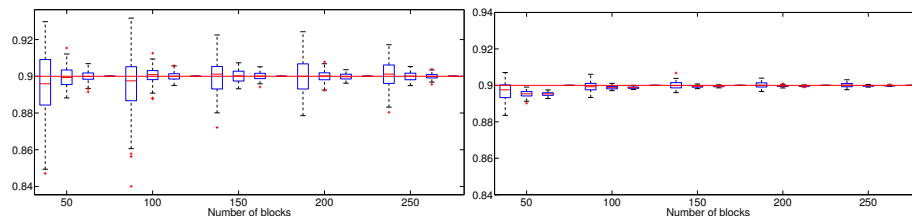
3.1 Linear Gaussian Model

Consider the Linear Gaussian model (LGM):

$$X_{t+1} = \phi X_t + \sigma_u U_t, \quad Y_t = X_t + \sigma_v V_t,$$

where $X_0 \sim \mathcal{N}(0, \sigma_u^2(1 - \phi^2)^{-1})$, $\{U_t\}_{t \geq 0}, \{V_t\}_{t \geq 0}$ are independent i.i.d. standard Gaussian r.v., independent from X_0 . Data are sampled using $\phi = 0.9$, $\sigma_u^2 = 0.6$ and $\sigma_v^2 = 1$. All runs are started with $\phi = 0.1$, $\sigma_u^2 = 1$ and $\sigma_v^2 = 2$.

We illustrate the convergence of the BOEM algorithms. We choose $\tau_n = a(n+1)$. We display in Figure 1 the box and whisker plots for the estimation of ϕ obtained with 100 independent Monte Carlo experiments; different values of a are also considered. Both BOEM and its averaged version converge to the true value $\phi = 0.9$; the averaging procedure clearly improves the variance of the estimation. Figure 1 shows that the averaged procedure needs a few more iterations to converge but when compared to the non averaged one, the variance is much smaller.



(a) BOEM without averaging, $\tau_n = a(n+1)$. (b) BOEM with averaging, $\tau_n = a(n+1)$.

Figure 1: Estimation of ϕ for $a = 10$ (left), $a = 100$ (middle) and $a = 300$ (right) after 50, 100, 150, 200 and 250 blocks.

We now discuss the role of the initial distribution χ . The convergence results (see Section 4) show that our algorithms converge whatever χ . Figure 2 displays

the estimation of ϕ by the averaged BOEM algorithm with $\tau_n \sim (n + 99)^{1.2}$, over 100 independent Monte Carlo runs as a function of the number of blocks. We consider first the case when χ is the stationary distribution of the hidden process, i.e. $\chi \equiv \mathcal{N}(0, (1 - \phi^2)^{-1} \sigma_u^2)$, computed with the current estimates, and the case when χ is the filtering distribution obtained at the end of the previous block, computed with the Kalman filter. In terms of the error of the estimation, the two strategies are similar. We observe the same phenomenon for different values of ϕ (plots are reported in [21, Section 5]). Therefore, it is advocated to choose χ as the filtering distribution obtained at the end of the previous block.

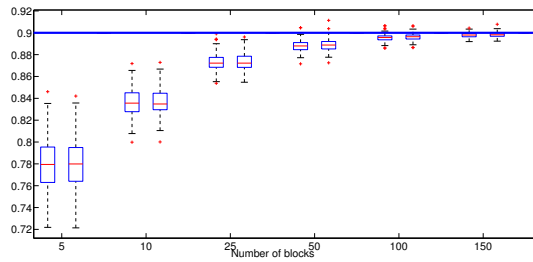
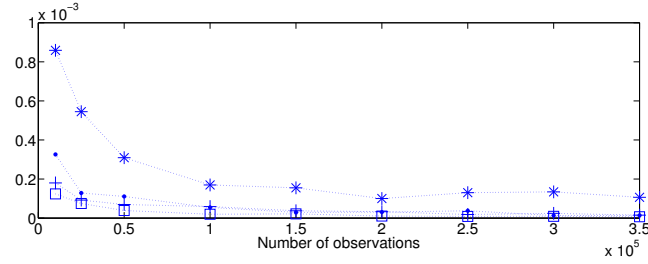


Figure 2: Estimation of ϕ after 5, 10, 25, 50, 100 and 150 blocks, with two different initialization schemes: the stationary distribution (left) and the filtering distribution at the end of the previous block (right). The boxplots are computed with 100 Monte Carlo runs.

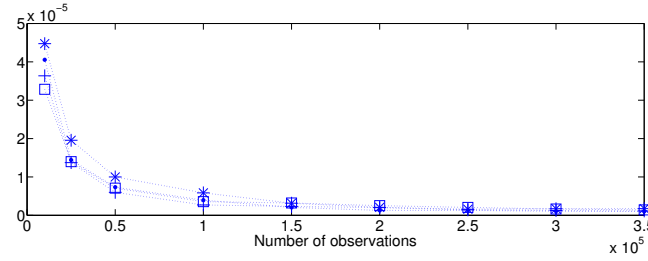
We now discuss the role of $\{\tau_n\}_{n \geq 0}$. Figure 3 displays the empirical variance, when estimating ϕ , computed with 100 independent Monte Carlo runs, for different numbers of observations and, for both the BOEM and its averaged version. We consider four polynomial rates $\tau_n \sim n^b$, $b \in \{1.2, 1.8, 2, 2.5\}$. Figure 3a shows that the choice of $\{\tau_n\}_{n \geq 0}$ has a great impact on the empirical variance of the (non averaged) BOEM path $\{\theta_n\}_{n \geq 0}$. To reduce this variability, a solution could consist in increasing the block sizes τ_n at a larger rate although this implies practical difficulties: when $\tau_n \sim n^2$, many observations are needed for each update of the parameter sequence. The influence of the block size sequence τ_n is greatly reduced with the averaging procedure as shown in Figure 3b. We will show in Section 5 that averaging really improves the rate of convergence of BOEM.

In addition, it is not advocated to start the averaging procedure with too few observations, as illustrated by Figure 4. The first estimates highly depend on the initialization of the parameter and the averaging procedure should start after a burn-in period.

As a conclusion, it is advocated to use the averaged BOEM algorithm. In practice, one could use slowly increasing sequences τ_n for the first iterations, and then, use more rapidly increasing sequences after the burn-in period.



(a) BOEM, without averaging



(b) BOEM, with averaging

Figure 3: BOEM: empirical variance of the estimation of ϕ after $n = 0.5\ell 10^5$ observations ($\ell \in \{1, \dots, 7\}$) for different block size schemes $\tau_n \sim n^{1.2}$ (stars), $\tau_n \sim n^{1.8}$ (dots), $\tau_n \sim n^2$ (crosses) and $\tau_n \sim n^{2.5}$ (squares).

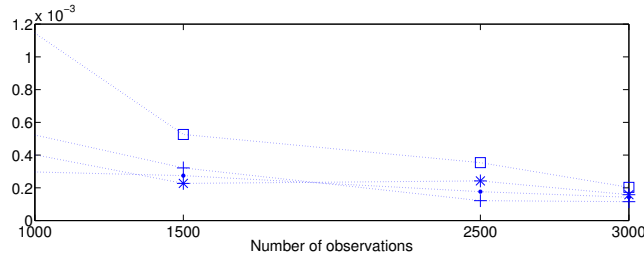


Figure 4: BOEM with averaging: empirical variance of the estimation of ϕ after $n = 1000, 1500, 2500$ and 3000 observations for different block size schemes $\tau_n \sim n^{1.2}$ (stars), $\tau_n \sim n^{1.8}$ (dots), $\tau_n \sim n^2$ (crosses) and $\tau_n \sim n^{2.5}$ (squares).

3.2 Finite state-space HMM

We consider models where the unobservable states take a finite number of values. Mixture processes with Markov dependence, switching processes with Markov regime, communication channels driven by Hidden Markov processes, composite sources with switch controlled by a Markov chain are examples of finite state-space HMM found useful in many fields including biostatistics, genomics, information theory, speech processing... (see e.g. [16] for a review). In the nu-

merical applications below, we consider a Gaussian mixture process with Markov dependence of the form: $Y_t = X_t + V_t$ where $\{X_t\}_{t \geq 0}$ is a Markov chain taking values in $\{x_1, \dots, x_d\}$, with initial distribution ν and a $d \times d$ transition matrix m . $\{V_t\}_{t \geq 0}$ are i.i.d. $\mathcal{N}(0, v)$ r.v., independent from $\{X_t\}_{t \geq 0}$. Observations are sampled using $d = 6$, $v = 0.5$, $x_i = i$, $\forall i \in \{1, \dots, d\}$ and the true transition matrix is given in [21, Section 5.2].

3.2.1 Comparison to an online EM based procedure

In this case, we want to estimate the variance v and the states $\{x_1, \dots, x_d\}$. All the runs are started from $v = 2$ and from the initial states $\{-1; 0; .5; 2; 3; 4\}$. We compare the averaged BOEM to the online EM procedure of [3] combined with a Polyak-Ruppert averaging (see [26]). The algorithm in [3] follows a stochastic approximation update and depends on a step-size sequence $\{\gamma_n\}_{n \geq 0}$. It is expected that the rate of convergence in L_2 after n observations is $\gamma_n^{1/2}$ (and $1/\sqrt{n}$ for its averaged version) - this assertion relies on classical results for stochastic approximation. We prove in Section 5 that the rate of convergence of BOEM is $n^{-b/(2(b+1))}$ (and $1/\sqrt{n}$ for its averaged version) when $\tau_n \propto n^b$. Therefore, we set $\tau_n = n^{1.1}$ and $\gamma_n = n^{-0.53}$. Figure 5 displays the empirical median and first and last quartiles for the estimation of v with both algorithms and their averaged versions as a function of the number of observations. These estimates are obtained over 100 independent Monte Carlo runs. Both BOEM and OEM converge to the true value of v and the averaged versions reduce the variability of the estimation. Figure 6 shows the similar behavior of both averaged algorithms for the estimation of x_1 in the same experiment. Nevertheless, while the online EM of [3] has an encouraging experimental behavior there is still no theoretical proof of convergence. Some supplementary graphs on the estimation of the states can be found in [21, Section 5]).

3.2.2 Comparison to a recursive maximum likelihood procedure

We want to estimate the variance v and the transition matrix m . All the runs are started from $v = 2$ and from a matrix m with each entry equal to $1/d$. The averaged BOEM is compared to a recursive maximum likelihood (RML) procedure (see [23, 28]) combined with Polyak-Ruppert averaging (see [26]). RML follows a stochastic approximation update and depends on a step-size sequence $\{\gamma_n\}_{n \geq 0}$ which is chosen in the same way as in Section 3.2.1. Therefore, for a fair comparison, RML (resp. BOEM) is run with $\gamma_n = n^{-0.53}$ (resp. $\tau_n = n^{1.1}$). Figure 7 displays the empirical median and empirical first and last quartiles of the estimation of $m(1, 1)$ as a function of the number of observations over 100 independent Monte Carlo runs. For both algorithms, the bias and the variance of the estimation decrease as n increases. Nevertheless, the bias and/or the variance of the averaged BOEM decrease faster than those of the averaged RML (similar graphs have been obtained for the estimation of the other entries of the matrix m and for the estimation of v ; see [21, Section 5]). As a conclusion, it is advocated to use the averaged BOEM instead of the averaged

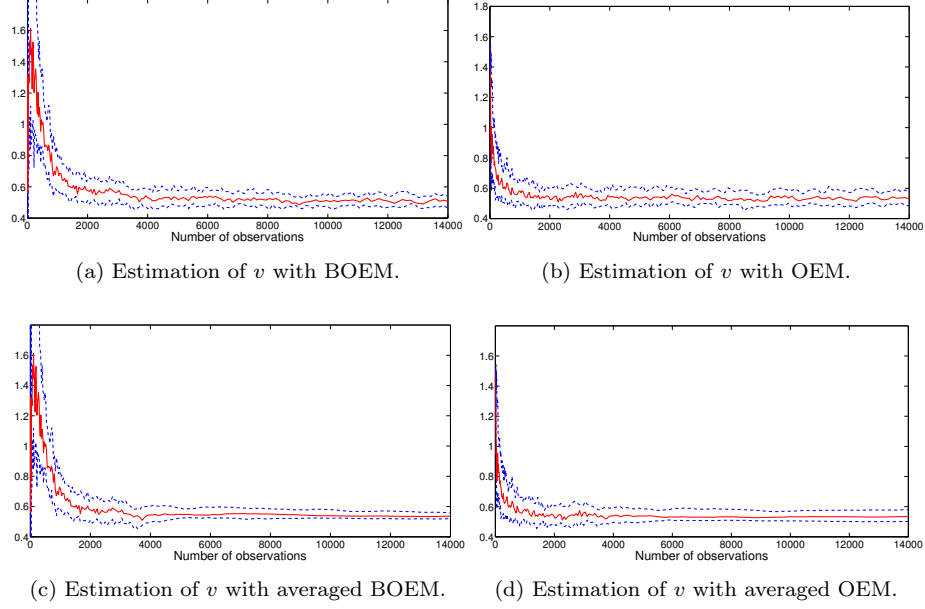


Figure 5: Estimation of v using the online EM and BOEM (top) and their averaged versions (bottom). Each plot displays the empirical median (bold line) and the first and last quartiles (dotted lines) over 100 independent Monte Carlo runs with $\tau_n = n^{1.1}$ and $\gamma_n = n^{-0.53}$.

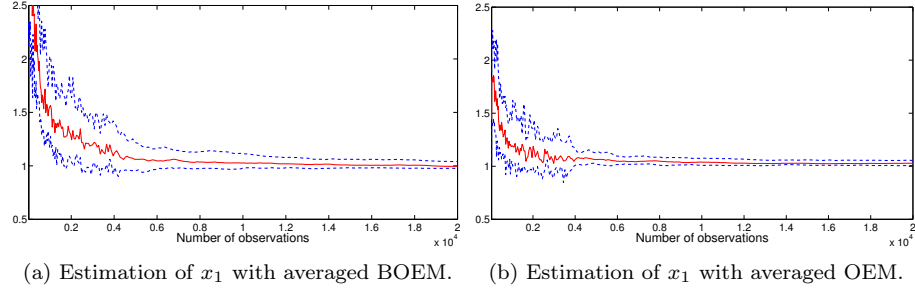


Figure 6: Estimation of x_1 using the averaged OEM and averaged BOEM. Each plot displays the empirical median (bold line) and the first and last quartiles (dotted lines) over 100 independent Monte Carlo runs with $\tau_n = n^{1.1}$ and $\gamma_n = n^{-0.53}$. The first ten observations are omitted for a better visibility.

RML.

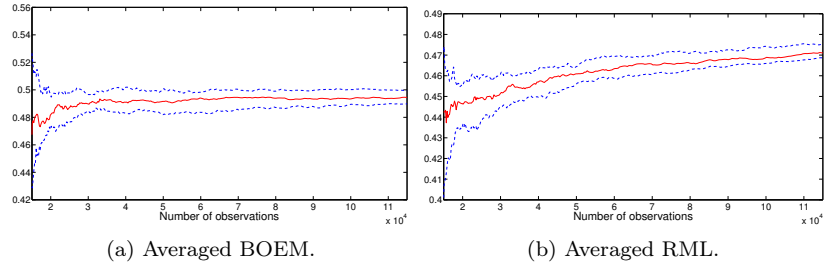


Figure 7: Empirical median (bold line) and first and last quartiles (dotted line) for the estimation of $m(1,1)$ using the averaged RML algorithm (right) and the averaged BOEM algorithm (left). The true values is $m(1,1) = 0.5$ and the averaging procedure is starter after 10000 observations. The first 10000 observations are not displayed for a better clarity.

4 Convergence of the Block Online EM algorithms

It is shown in Section 4.2 that for any $T > 0$ and any initial distribution χ , the quantity $\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y})$ converges \mathbb{P}_\star -a.s when $\tau \rightarrow +\infty$, to a deterministic quantity $\bar{S}(\theta)$ that does not depend on T and χ . Therefore, the BOEM algorithm can be seen as a perturbation of the so-called *limiting EM* algorithm, defined as a deterministic iterative algorithm $\bar{\theta}_n = R(\bar{\theta}_{n-1})$ where

$$R(\theta) \stackrel{\text{def}}{=} \bar{\theta}(\bar{S}(\theta)) . \quad (9)$$

The limiting points of the *limiting EM* algorithm are identified (see section 4.3) and it is shown in Section 4.4 that BOEM inherits this limiting behavior provided the perturbation can be set small enough. All the convergence results are addressed under the assumptions introduced in Section 4.1.

4.1 Assumptions

A2 There exist σ_- and σ_+ s.t. for any $(x, x') \in \mathbb{X}^2$ and any $\theta \in \Theta$, $0 < \sigma_- \leq m_\theta(x, x') \leq \sigma_+$. Set $\rho \stackrel{\text{def}}{=} 1 - (\sigma_-/\sigma_+)$.

This assumption is known in the literature as the *strong mixing condition*. It is commonly used to prove the forgetting property of the initial condition of the filter, see e.g. [9, 10]. This assumption holds for example if \mathbb{X} is finite and for any $(x, x') \in \mathbb{X}^2$, $0 < \inf_\theta m_\theta(x, x') \leq \sup_\theta m_\theta(x, x') < +\infty$. Under regularity conditions on the kernels $\{m_\theta; \theta \in \Theta\}$, it also holds when \mathbb{X} is compact. Nevertheless, it fails to hold in standard situations s.t. linear and Gaussian state-space models. It has been weakened in recent works: in [13], the exponential forgetting of the initial condition of the filter is proved with a local Doeblin property; [30] gives an uniform time average convergence of some particle filters. The approach in [13] could be adapted to the present paper but at a quite

technical cost. For pedagogical purposes, we will assume A2 throughout this paper.

We now introduce assumptions on the observation process. Define the shift operator ϑ onto $\mathbb{Y}^{\mathbb{Z}}$ by $(\vartheta \mathbf{y})_k = \mathbf{y}_{k+1}$ for any $k \in \mathbb{Z}$; and by induction, define the s -iterated shift operator $\vartheta^{s+1} \mathbf{y} = \vartheta(\vartheta^s \mathbf{y})$, with the convention that ϑ^0 is the identity operator. The shift operator is said to be ergodic for \mathbb{P}_\star if for each set A in $\{A \in \mathcal{B}(\mathbb{Y})^{\otimes \mathbb{Z}}; A = \vartheta^{-1}(A)\}$, $\mathbb{P}_\star(A) \in \{0, 1\}$ (see [1, p.314]).

A3-(γ) $\mathbb{E}_\star [\sup_{x, x' \in \mathbb{X}^2} |S(x, x', Y_0)|^\gamma] < +\infty$.

A4 (a) Under \mathbb{P}_\star , \mathbf{Y} is a stationary sequence.

(b) The shift operator is ergodic with respect to \mathbb{P}_\star .

(c) $\mathbb{E}_\star [|\log b_-(Y_0)| + |\log b_+(Y_0)|] < +\infty$ where

$$b_-(y) \stackrel{\text{def}}{=} \inf_{\theta \in \Theta} \int g_\theta(x, y) \lambda(dx), \quad b_+(y) \stackrel{\text{def}}{=} \sup_{\theta \in \Theta} \int g_\theta(x, y) \lambda(dx).$$

Finally, assumptions on the forgetting properties of the observations \mathbf{Y} are required. For any sequence of r.v. $Z \stackrel{\text{def}}{=} \{Z_t\}_{t \in \mathbb{Z}}$ on $(\Omega, \tilde{\mathbb{P}}, \mathcal{F})$, let

$$\mathcal{F}_k^Z \stackrel{\text{def}}{=} \sigma(\{Z_u\}_{u \leq k}) \quad \text{and} \quad \mathcal{G}_k^Z \stackrel{\text{def}}{=} \sigma(\{Z_u\}_{u \geq k}) \quad (10)$$

be σ -fields associated to Z . We also define the mixing coefficients by, see [7],

$$\beta^Z(n) = \sup_{u \in \mathbb{Z}} \sup_{B \in \mathcal{G}_{u+n}^Z} |\tilde{\mathbb{P}}(B | \mathcal{F}_u^Z) - \tilde{\mathbb{P}}(B)|, \quad \forall n \geq 0. \quad (11)$$

A5 There exist $C \in [0, 1)$ and $\beta \in (0, 1)$ s.t. for any $n \geq 0$, $\beta^{\mathbf{Y}}(n) \leq C\beta^n$, where $\beta^{\mathbf{Y}}$ is defined in (11).

Under A4(a), the shift operator preserves the measure \mathbb{P}_\star on $(\mathbb{Y}^{\mathbb{Z}}, \mathcal{B}(\mathbb{Y})^{\otimes \mathbb{Z}})$. A5 is used to control the L_p -mean error between the *limiting EM* map $\bar{\theta}(\bar{S}(\theta))$ and a BOEM iteration $\bar{\theta}(S_{\tau_n}^{X, T_{n-1}}(\theta, \mathbf{Y}))$ both started from the same point θ . Examples of observation processes satisfying A4(b) and A5 include geometrically ergodic Markov chains as discussed in [21, Section 2.1].

We conclude this set of assumptions by a condition on the block size sequence.

A6-(γ) The block size sequence $\{\tau_n\}_{n \geq 1}$ satisfies $\sum_{k \geq 0} \tau_k^{-\gamma/2} < \infty$.

4.2 Block Online EM and *limiting* EM algorithms

Theorem 4.1. Assume A2 and A4(a-b). Let $S : \mathbb{X}^2 \times \mathbb{Y} \rightarrow \mathbb{R}^d$ be a measurable function s.t. A3-(1) holds. For any $\theta \in \Theta$, there exists a \mathbb{P}_\star -integrable

r.v. denoted by $\mathbb{E}_\theta [S(X_{-1}, X_0, Y_0)|\mathbf{Y}]$ s.t. for any probability distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$,

$$\begin{aligned} \sup_{\theta \in \Theta} \left| \Phi_{\theta, 0, \tau}^{\chi, -\tau}(S, \mathbf{Y}) - \mathbb{E}_\theta [S(X_{-1}, X_0, Y_0)|\mathbf{Y}] \right| \\ \leq 2(\rho^\tau + \rho^{\tau-1}) \sup_{(x, x') \in \mathbb{X}^2} |S(x, x', Y_0)| \quad \mathbb{P}_\star - a.s. \end{aligned} \quad (12)$$

Define for all $\theta \in \Theta$,

$$\bar{S}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_\star [\mathbb{E}_\theta [S(X_{-1}, X_0, Y_0)|\mathbf{Y}]] . \quad (13)$$

$\theta \mapsto \bar{S}(\theta)$ is continuous on Θ and for any $T > 0$,

$$\bar{S}_\tau^{\chi, T}(\theta, \mathbf{Y}) \xrightarrow{\tau \rightarrow +\infty} \bar{S}(\theta) \quad \mathbb{P}_\star - a.s. \quad (14)$$

The proof of Theorem 4.1 is given in Section 6.1. Eqs (2) and (12) show that when $\{(X_t, Y_t)\}_{t \in \mathbb{Z}}$ is a HMM with transition kernels m_θ and g_θ , the limiting statistic $\mathbb{E}_\theta [S(X_{-1}, X_0, Y_0)|\mathbf{Y}]$ is the a.s. limit of the conditional expectation of $S(X_{-1}, X_0, Y_0)$ given $Y_{-\tau+1:\tau}$ when $X_{-\tau} \sim \chi$, whatever χ is:

$$\mathbb{E}_\theta [S(X_{-1}, X_0, Y_0)|Y_{-\tau+1:\tau}] \xrightarrow{\tau \rightarrow +\infty} \mathbb{E}_\theta [S(X_{-1}, X_0, Y_0)|\mathbf{Y}] \quad \mathbb{P}_\star - a.s. .$$

\bar{S} is the \mathbb{P}_\star -a.s. limit of the usual sufficient statistics of the EM algorithm when the number of observations grows to infinity. Hence, the *limiting EM* can be seen as an EM algorithm with the whole data set \mathbf{Y} : since \mathbf{Y} is stationary, for this *limiting EM*, the so-called sufficient statistics (in exponential HMM) depend on the observations only through the mean $\mathbb{E}_\star [\mathbb{E}_\theta [S(X_{-1}, X_0, Y_0)|\mathbf{Y}]]$.

As a consequence of (14), when τ is large, the quantity $\bar{S}_\tau^{\chi, T}(\theta, \mathbf{Y})$ is an approximation of $\bar{S}(\theta)$. Therefore, the BOEM algorithm (6) is a perturbation of the *limiting EM* algorithm (9). Based on this remark, we first address the convergence of the *limiting EM* and then we show that BOEM has the same behavior.

4.3 Asymptotic behavior of the *limiting EM*

The convergence of the *limiting EM* is addressed following the same approach as in [32] for the convergence of the EM algorithm. It relies on a Lyapunov function W w.r.t. to the map R and the set

$$\mathcal{L} \stackrel{\text{def}}{=} \{\theta \in \Theta; R(\theta) = \theta\} . \quad (15)$$

The existence of such a Lyapunov function is the key ingredient to identify the limiting points of the algorithm (9).

Proposition 4.2. *Assume A1-2, A3-(1) and A4. Then R given by (9) is continuous on Θ and there exists a continuous function on Θ , W , s.t.*

(i) For all $\theta \in \Theta$, $W \circ R(\theta) - W(\theta) \geq 0$.

(ii) For all compact set $\mathcal{K} \subset \Theta \setminus \mathcal{L}$, $\inf_{\theta \in \mathcal{K}} \{W \circ R(\theta) - W(\theta)\} > 0$.

Proposition 4.2 is proved in Section 6.2. The following proposition gives a set of sufficient conditions for the convergence of the *limiting EM* algorithm $\check{\theta}_n = R(\check{\theta}_{n-1})$ to the set \mathcal{L} (see [17, Proposition 9] for the proof).

Proposition 4.3. *Assume A1-2, A3-(1) and A4. Assume in addition that for any $M > 0$, the set $\mathcal{K}_M \stackrel{\text{def}}{=} \{\theta \in \Theta; W(\theta) \geq M\}$ is a compact subset of Θ . Then, for any initial value $\check{\theta}_0$ s.t. $W(\mathcal{K}_{W(\check{\theta}_0)} \cap \mathcal{L})$ has an empty interior, there exists w_\star s.t. $\{\check{\theta}_k\}_{k \geq 0}$ converges to $\{\theta \in \mathcal{L}; W(\theta) = w_\star\}$.*

It is well known that for EM, a natural Lyapunov function is based on the (normalized) log-likelihood of the observations (see e.g. [32]). [14, Lemma 2 and Proposition 1] shows that, under A2-3, the normalized log-likelihood converges and this limit, hereafter denoted by $c_\star(\theta)$, is deterministic and does not depend on the initial distribution χ of the hidden chain. To prove Proposition 4.2, we establish that the function $W : \theta \mapsto \exp(c_\star(\theta))$ is a Lyapunov function for the map R and the set \mathcal{L} . It can be proved that under regularity conditions on the HMM, the set \mathcal{L} is the set of the stationary points of c_\star ; this discussion is detailed in [21, Theorem 14]. By Sard's theorem if W is at least d_θ (where $\Theta \subset \mathbb{R}^{d_\theta}$) continuously differentiable, then $W(\mathcal{L})$ has Lebesgue measure 0 and hence has an empty interior.

The assumptions on the compacity of the level sets \mathcal{K}_M highly depend on the model. In [32], the same assumption is used to prove that the limit points of the EM algorithm are the stationary points of the likelihood of the observations. In [11, 17], the stability of stochastic-type EM algorithms rely on these assumptions. Since W is continuous, the compacity of the level set can be proved if $\lim_{\theta \rightarrow \partial\Theta} W(\theta) = 0$, as $\theta \rightarrow \partial\Theta$.

4.4 Asymptotic behavior of the Block Online EM algorithms

Theorem 4.4 establishes the convergence of BOEM. Let $\text{Cl}(A)$ be the closure of the set A .

Theorem 4.4. *Assume A1-2, A3-(\bar{p}_2), A4-5 and A6-(\bar{p}_1) for some $2 < \bar{p}_1 < \bar{p}_2$. Assume in addition that $W(\mathcal{L})$ is compact and, for any $M > 0$, the level set $\{\theta \in \Theta; W(\theta) \geq M\}$ is compact. Then,*

(a) $\limsup_n p_n < +\infty$ \mathbb{P}_\star -a.s where p_n is defined in (6).

(b) If $W(\mathcal{L} \cap \text{Cl}(\{\theta_n\}_{n \geq 0}))$ has an empty interior, there exists w_\star s.t. $\{\theta_n\}_{n \geq 0}$ converges to $\{\theta \in \mathcal{L}; W(\theta) = w_\star\}$.

Theorem 4.4 implies that the number of truncations p_n in (6) is almost surely finite so that for a (random) sufficiently large n , $\theta_n = \theta_{n-1/2}$. It shows

that the BOEM algorithm and the *limiting EM* have the same asymptotic behavior. The convergence of $\{\theta_n\}_{n \geq 0}$ is established under the same assumptions on $W(\mathcal{L} \cap \text{Cl}(\{\theta_n\}_{n \geq 0}))$ as in Proposition 4.3 for the convergence of $\{\tilde{\theta}_n\}_{n \geq 0}$ (see above for comments on this assumption). The proof is detailed in Section 6.3: it consists in applying the results of [17] on the convergence of a sequence generated by iterated random maps, which are perturbations of a point-to-point map associated to a Lyapunov function. The key ingredient is to prove that the perturbation vanishes when the number of iterations tends to infinity; in our case, this is done through the control of the L_p -mean error when replacing the limiting quantity $\bar{S}(\theta_{n-1})$ by $\bar{S}_{\tau_n}^{X, T_{n-1}}(\theta_{n-1}, \mathbf{Y})$ (see Proposition 6.5 in Section 6.3).

Then, we show that $\{W(\theta_n)\}_{n \geq 0}$ converges to a connected component of $W(\mathcal{L})$ and when $W(\mathcal{L} \cap \text{Cl}(\{\theta_n\}_{n \geq 0}))$ has an empty interior $\{W(\theta_n)\}_{n \geq 0}$ converges to a point ω_* . We then deduce the convergence of $\{\theta_n\}_{n \geq 0}$.

A convergence result for the averaged BOEM algorithm can be obtained following the same lines as in the proof of Theorem 4.4. The main ingredient for this proof is the control of the L_p -mean error when replacing $\bar{S}(\theta_{n-1})$ by Σ_n (see Lemma 6.7 below). It can be proved that, along any converging BOEM path, the averaged BOEM algorithm and the *limiting EM* have the same asymptotic behavior. Details are omitted for brevity.

5 Rate of convergence of the Block Online EM algorithm

We address the rate of convergence of $\{\theta_n\}_{n \geq 0}$ and $\{\tilde{\theta}_n\}_{n \geq 0}$, resp. given by (6) and (8) to a point $\theta_* \in \mathcal{L}$ (see Theorem 4.4). It is assumed that

- A7** (a) \bar{S} and $\bar{\theta}$ are twice continuously differentiable on Θ and \mathcal{S} .
(b) There exists $0 < \gamma < 1$ s.t. $\text{sp}(\nabla_s(\bar{S} \circ \bar{\theta})_{s=\bar{S}(\theta_*)}) \leq \gamma$ where sp denotes the spectral norm.

- A8** (a) $\{\tau_{n+1}/\tau_n\}_{n \geq 0}$ converges to q and $\gamma q < 1$.

(b) $\limsup_n \sum_{k=1}^n \left\{ \left| \frac{\tau_{k+1}}{\tau_k} - q \right| \sqrt{\tau_k} + \log \tau_k \right\} / \sqrt{T_n} < \infty$.

Under A6, $\lim_n \tau_n = +\infty$. A8 strengthens A6. A8(a) is satisfied for geometric rates of the form $\tau_n \sim a\tau^n$ with $\tau \in (1, \gamma^{-1})$, for polynomial rates $\tau_n \sim cn^b$ with $b > 0$ and sub-exponential rates $\log \tau_n \sim cn^b$ with $c > 0, b \in (0, 1)$, and more generally with sub-geometric rates. A8(b) is satisfied for geometric rates of the form $\tau_n \sim a\tau^n$ with $\tau > 1$, for polynomial rates of the form $\tau_n \sim cn^b$ with $b \geq 1$ and with any sub-exponential rates.

Hereafter, for any sequence of random variables $\{Z_n\}_{n \geq 0}$, write $Z_n = \mathcal{O}_{L_p}(1)$ if $\limsup_n \mathbb{E}_* [|Z_n|^p] < \infty$; $Z_n = \mathcal{O}_{\text{a.s.}}(1)$ if $\sup_n |Z_n| < +\infty$ $\mathbb{P}_* - \text{a.s.}$ and $Z_n = o_{\text{a.s.}}(1)$ if $\lim_{n \rightarrow +\infty} |Z_n| = 0$ $\mathbb{P}_* - \text{a.s.}$

Theorem 5.1. Assume A2, A3-(\bar{p}_2), A4-5, A6-(\bar{p}_1), A7 and A8(a) for some $2 < \bar{p}_1 < \bar{p}_2$. Then, for any $p \in (2, \bar{p}_2)$,

$$\sqrt{\tau_n} [\theta_n - \theta_\star] \mathbf{1}_{\lim_n \theta_n = \theta_\star} = \mathcal{O}_{L_p}(1) + \frac{1}{\sqrt{\tau_n}} \mathcal{O}_{L_{p/2}}(1) \mathcal{O}_{a.s.}(1) + o_{a.s.}(1) . \quad (16)$$

If in addition A8(b) holds, then for any $p \in (2, \bar{p}_2)$,

$$\sqrt{T_n} [\tilde{\theta}_n - \theta_\star] \mathbf{1}_{\lim_n \theta_n = \theta_\star} = \mathcal{O}_{L_p}(1) + \frac{n}{\sqrt{T_n}} \mathcal{O}_{L_{p/2}}(1) \mathcal{O}_{a.s.}(1) . \quad (17)$$

The proof of Theorem 5.1 is given in Section 6.4. Eq. (16) shows that the error $\theta_n - \theta_\star$ is decomposed into two terms and the L_p -norm of the leading term is inversely proportional to $\tau_n^{1/2}$. Hence, the rate of BOEM is closely related to the choice of the number of observations per block. The first column of Table 1 gives explicit rates of convergence for different block-sizes.

In (16), the rate is a function of the number of updates (i.e. the number of iteration of the algorithm). This rate could also be interpreted as a function of the total number of observations up to iteration n . To that goal, let $\phi(n) + 1$ be the index of the block the n -th observation belongs to, i.e. $\phi(n)$ is the largest integer s.t.

$$\sum_{k=0}^{\phi(n)} \tau_k < n \leq \sum_{k=0}^{\phi(n)+1} \tau_k , \text{ (by convention, } \sum_{k=0}^{-1} \tau_k = 0) .$$

The interpolated sequence $\{\theta_n^i\}_{n \geq 0}$ deduced from $\{\theta_n\}_{n \geq 0}$ is thus defined by $\theta_n^i = \theta_{\phi(n)}$ (the value of the interpolated sequence is kept fixed within each block). The second column of Table 1 gives the rate of convergence of this interpolated sequence (deduced from the square root of $\tau_{\phi(n)}$) up to a multiplicative constant. This rate of convergence is slower than $n^{-1/2}$, except in the geometric case. Note however that the geometric case is of weak practical interest, since the parameter is hardly ever updated thus yielding to algorithms which are really sensible to the initial value θ_0 (see Section 3).

Eq. (17) addresses the rate of convergence of the averaged BOEM algorithm. It shows that when the condition A8 is strengthened in such a way that $\lim_n n/\sqrt{T_n} = 0$, averaging reduces the influence of the block-size schedule: the error $\tilde{\theta}_n - \theta_\star$ has a rate of convergence proportional to $T_n^{-1/2}$ i.e. to the inverse of the square root of the total number of observations up to iteration n . The last column of Table 1 shows that this averaging procedure gives an optimal rate of convergence, whatever the block-size sequence.

As a conclusion, the averaged BOEM algorithm reaches the optimal rate of convergence even when the block size sequence $\{\tau_n\}_{n \geq 0}$ slowly increases, thus allowing polynomially increasing size of blocks.

τ_n	$\tau_n^{1/2}$	$\tau_{\phi(n)}^{1/2}$	$T_n^{1/2}$	$T_{\phi(n)}^{1/2}$
$c n^b, (b > 1)$	$n^{b/2}$	$n^{b/(2(b+1))}$	$n^{(b+1)/2}$	$n^{1/2}$
$\exp(c n^b), (b \in (0, 1))$	$\exp(0.5 c n^b)$	$n^{1/2} (\ln n)^{(b-1)/(2b)}$	$n^{(1-b)/2} \exp(0.5 c n^b)$	$n^{1/2}$
$c \tau^n, (\tau \in (1, \gamma^{-1}))$	$\tau^{n/2}$	$n^{1/2}$	$\tau^{n/2}$	$n^{1/2}$

Table 1: Rate of convergence of both algorithms (up to a multiplicative constant)

6 Proofs

For $p > 0$ and Z a random variable measurable w.r.t. the σ -algebra $\sigma(Y_n, n \in \mathbb{Z})$, set $\|Z\|_{*,p} \stackrel{\text{def}}{=} (\mathbb{E}_* [|Z|^p])^{1/p}$.

6.1 Proof of Theorem 4.1

The proof of Theorem 4.1 relies on auxiliary results about the forgetting properties of HMM. Most of them are really close to published results and their proof is provided in the supplementary material [21, Section 4]. The main novelty is the forgetting property of the bivariate smoothing distribution.

Lemma 6.1. *Assume A1-2. Let $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ s.t. $\sup_{x,x'} |S(x, x', y_i)| < +\infty$ for any $i \in \mathbb{Z}$. Then for any $r > 0$ and any distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, $\theta \mapsto \Phi_{\theta,0,r}^{\chi,-r}(S, \mathbf{y})$ is continuous on Θ .*

Proof. Set $K_{\theta}(x, x', y) \stackrel{\text{def}}{=} m_{\theta}(x, x') g_{\theta}(x', y)$. Let $r > 0$ and χ be a distribution on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. By definition of $\Phi_{\theta,0,r}^{\chi,-r}(S, \mathbf{y})$ (see (1)) we have to prove that

$$\theta \mapsto \int \chi(dx_{-r}) \left(\prod_{i=-r}^{r-1} K_{\theta}(x_i, x_{i+1}, y_{i+1}) \right) h(x_{-1}, x_0, y_0) \, d\lambda(x_{-r+1:r})$$

is continuous for $h(x, x', y) = 1$ and $h(x, x', y) = S(x, x', y)$. By A1(a), the function $\theta \mapsto \prod_{i=-r}^{r-1} K_{\theta}(x_i, x_{i+1}, y_{i+1}) h(x_{-1}, x_0, y_0)$ is continuous. In addition, under A1, for any $\theta \in \Theta$,

$$\begin{aligned} & \left| \prod_{i=-r}^{r-1} K_{\theta}(x_i, x_{i+1}, y_{i+1}) h(x_{-1}, x_0, y_0) \right| \\ &= |h(x_{-1}, x_0, y_0)| \exp \left(2r\phi(\theta) + \left\langle \psi(\theta), \sum_{i=-r}^{r-1} S(x_i, x_{i+1}, y_{i+1}) \right\rangle \right). \end{aligned}$$

Let \mathcal{K} be a compact subset of Θ . By A1, there exist constants C_1 and C_2 s.t. the supremum in $\theta \in \Theta$ of this expression is bounded above by

$$C_1 \sup_{x,x'} |h(x, x', y_0)| \exp \left(C_2 \sum_{i=-r}^{r-1} \sup_{x,x'} |S(x, x', y_{i+1})| \right).$$

Since χ is a distribution and λ is a finite measure, the continuity follows from the dominated convergence theorem. \square

Let us introduce the following shorthand $S_s(x, x') \stackrel{\text{def}}{=} S(x, x', Y_s)$. For a function h , define $\text{osc}(h) \stackrel{\text{def}}{=} \sup_{z, z'} |h(z) - h(z')|$. Note that under A3-(1), $\mathbb{E}_\star[\text{osc}(S_0)] < +\infty$. Under A2, [21, Proposition 4.3(ii)] implies that for any $\theta \in \Theta$, there exists a r.v. $\Phi_\theta(S, \mathbf{Y})$ s.t. for any $r < s \leq T$,

$$\sup_{\theta \in \Theta} \left| \Phi_{\theta, s, T}^{\chi, r}(S, \mathbf{Y}) - \Phi_\theta(S, \vartheta^s \mathbf{Y}) \right| \leq (\rho^{T-s} + \rho^{s-r-1}) \text{osc}(S_s). \quad (18)$$

This concludes the proof of (12). For the proof of (14), we introduce the following decomposition: for all $T > 0$,

$$\bar{S}_\tau^{\chi, T}(\theta, \mathbf{Y}) = \frac{1}{\tau} \sum_{t=1}^{\tau} \left[\Phi_\theta(S, \vartheta^{t+T} \mathbf{Y}) + \left\{ \Phi_{\theta, t, \tau}^{\chi, 0}(S, \vartheta^T \mathbf{Y}) - \Phi_\theta(S, \vartheta^{t+T} \mathbf{Y}) \right\} \right],$$

upon noting that by (3), $\bar{S}_\tau^{\chi, T}(\theta, \mathbf{Y}) = \tau^{-1} \sum_{t=1}^{\tau} \Phi_{\theta, t, \tau}^{\chi, 0}(S, \vartheta^T \mathbf{Y})$. By (1), (18) and A3-(1) $\mathbb{E}_\star[|\Phi_\theta(S, \mathbf{Y})|] < +\infty$. Under A4(a-b), the ergodic theorem (see e.g. [1, Theorem 24.1, p.314]) states that

$$\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \Phi_\theta(S, \vartheta^{t+T} \mathbf{Y}) = \mathbb{E}_\star[\Phi_\theta(S, \mathbf{Y})] \quad \mathbb{P}_\star - \text{a.s.}$$

for any fixed T . By (18),

$$\frac{1}{\tau} \sum_{t=1}^{\tau} \left| \Phi_{\theta, t, \tau}^{\chi, 0}(S, \vartheta^T \mathbf{Y}) - \Phi_\theta(S, \vartheta^{t+T} \mathbf{Y}) \right| \leq \frac{1}{\tau} \sum_{t=1}^{\tau} (\rho^{\tau-t} + \rho^{t-1}) \text{osc}(S_{t+T}). \quad (19)$$

Set $Z_t \stackrel{\text{def}}{=} \frac{1}{t} \sum_{s=1}^t \text{osc}(S_{s+T})$ and $Z_0 \stackrel{\text{def}}{=} 0$. Then, by an Abel transform,

$$\frac{1}{\tau} \sum_{t=1}^{\tau} \rho^{t-1} \text{osc}(S_{t+T}) = \rho^{\tau-1} Z_\tau + \frac{1-\rho}{\tau} \sum_{t=1}^{\tau-1} t \rho^{t-1} Z_t. \quad (20)$$

Under A4(a-b) and A3-(1), the ergodic theorem implies that $\lim_{\tau \rightarrow \infty} Z_\tau = \mathbb{E}_\star[\text{osc}(S_0)] \mathbb{P}_\star - \text{a.s.}$ Therefore, $\limsup_{\tau} Z_\tau < \infty \mathbb{P}_\star - \text{a.s.}$ Since $\sum_{t \geq 1} t \rho^{t-1} < \infty$, this implies that $\tau^{-1} \sum_{t=1}^{\tau} \rho^{t-1} \text{osc}(S_{t+T}) \xrightarrow{\tau \rightarrow +\infty} 0 \mathbb{P}_\star - \text{a.s.}$ Similarly,

$$\frac{1}{\tau} \sum_{t=1}^{\tau} \rho^{\tau-t} \text{osc}(S_{t+T}) = Z_\tau - (1-\rho) \sum_{t=1}^{\tau-1} \rho^{\tau-t-1} Z_t + \frac{1-\rho}{\tau} \sum_{t=1}^{\tau-1} t \rho^{t-1} Z_{\tau-t}.$$

We have $\lim_{\tau \rightarrow \infty} \tau^{-1} \sum_{t=1}^{\tau-1} t \rho^{t-1} Z_{\tau-t} = 0, \mathbb{P}_\star - \text{a.s.}$ by using the same arguments as for the second term in (20). Furthermore,

$$\left| \sum_{t=1}^{\tau-1} \frac{\rho^{\tau-t-1}}{1-\rho} Z_t - \mathbb{E}_\star[\text{osc}(S_0)] \right| \leq \sum_{t=1}^{\tau-1} \frac{\rho^{\tau-t-1}}{1-\rho} |Z_t - \mathbb{E}_\star[\text{osc}(S_0)]| + \mathbb{E}_\star[\text{osc}(S_0)] \rho^{\tau-1}.$$

Since $Z_\tau \xrightarrow{\tau \rightarrow +\infty} \mathbb{E}_\star [\text{osc}(S_0)] \mathbb{P}_\star - \text{a.s.}$, the RHS converges $\mathbb{P}_\star - \text{a.s.}$ to 0 and

$$\lim_{\tau \rightarrow +\infty} \left| Z_\tau - (1 - \rho) \sum_{t=1}^{\tau-1} \rho^{\tau-t-1} Z_t \right| = 0 \quad \mathbb{P}_\star - \text{a.s.}$$

Hence, the RHS in (19) converges $\mathbb{P}_\star - \text{a.s.}$ to 0 and this concludes the proof of (14). We now prove that the function $\theta \mapsto \mathbb{E}_\star [\Phi_\theta(S, \mathbf{Y})]$ is continuous by application of the dominated convergence theorem. From [21, Proposition 4.3(ii)], for any \mathbf{y} s.t. $\text{osc}(S(\cdot, \cdot, y_0)) < \infty$,

$$\lim_{r \rightarrow +\infty} \sup_{\theta \in \Theta} \left| \Phi_{\theta,0,r}^{\chi, -r}(S, \mathbf{y}) - \Phi_\theta(S, \mathbf{y}) \right| = 0.$$

Then, by Lemma 6.1, $\theta \mapsto \Phi_\theta(S, \mathbf{y})$ is continuous for any \mathbf{y} such that $\text{osc}(S(\cdot, \cdot, y_0)) < +\infty$. In addition, $\sup_{\theta \in \Theta} |\Phi_\theta(S, \mathbf{Y})| \leq \sup_{x, x'} |S(x, x', Y_0)|$. We then conclude by A3-(1).

6.2 Proof of Proposition 4.2

Set

$$\ell_{\theta,T}^{\chi,0}(\mathbf{Y}) \stackrel{\text{def}}{=} \log \left(\int \chi(dx_0) \left\{ \prod_{t=1}^T m_\theta(x_{t-1}, x_t) g_\theta(x_t, Y_t) \right\} \lambda(dx_1) \cdots \lambda(dx_T) \right).$$

(Continuity of R and W) By A1(c) and Theorem 4.1, the function R is continuous. Under A1-2 and A4, there exists a continuous function c_\star on Θ s.t. $\lim_T T^{-1} \ell_{\theta,T}^{\chi,0}(\mathbf{Y}) = c_\star(\theta) \mathbb{P}_\star - \text{a.s.}$ for any distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and any $\theta \in \Theta$, (see [14, Lemma 2 and Proposition 1], see also [21, Theorem 4.9]). Therefore, W is continuous.

Proof of Proposition 4.2 (i) For all $T > 0$ and all $\theta \in \Theta$, define

$$p_\theta(x_{0:T}, Y_{1:T}) \stackrel{\text{def}}{=} \prod_{i=1}^T m_\theta(x_{i-1}, x_i) g_\theta(x_i, Y_i). \quad (21)$$

Under Assumption A1(a)

$$\frac{1}{T} \log p_\theta(x_{0:T}, Y_{1:T}) = \phi(\theta) + \left\langle \left\{ \frac{1}{T} \sum_{t=1}^T S(x_{t-1}, x_t, \mathbf{Y}_t) \right\}, \psi(\theta) \right\rangle.$$

Upon noting that

$$\int S(x_{t-1}, x_t, Y_t) \frac{p_\theta(x_{0:T}, Y_{1:T})}{\int p_\theta(z_{0:T}, Y_{1:T}) \lambda(dz_{1:T}) \chi(dz_0)} \lambda(dx_{1:T}) \chi(dx_0) = \Phi_{\theta,t,T}^{\chi,0}(S, \mathbf{Y}),$$

a classical use of the Jensen inequality gives, \mathbb{P}_\star - a.s.,

$$\begin{aligned} \frac{1}{T} \ell_{\mathbf{R}(\theta), T}^{\chi, 0}(\mathbf{Y}) - \frac{1}{T} \ell_{\theta, T}^{\chi, 0}(\mathbf{Y}) &\geq \phi(\mathbf{R}(\theta)) + \left\langle \frac{1}{T} \sum_{t=1}^T \Phi_{\theta, t, T}^{\chi, 0}(S, \mathbf{Y}), \psi(\mathbf{R}(\theta)) \right\rangle \\ &\quad - \phi(\theta) - \left\langle \frac{1}{T} \sum_{t=1}^T \Phi_{\theta, t, T}^{\chi, 0}(S, \mathbf{Y}), \psi(\theta) \right\rangle. \end{aligned} \quad (22)$$

Under A1-4, it holds by Theorem 4.1 and [14, Lemma 2 and Proposition 1] (see also [21, Theorem 4.9(ii)]) that for all $\theta \in \Theta$, \mathbb{P}_\star - a.s.,

$$\frac{1}{T} \sum_{t=1}^T \Phi_{\theta, t, T}^{\chi, 0}(S, \mathbf{Y}) \xrightarrow{T \rightarrow +\infty} \bar{S}(\theta), \quad \frac{1}{T} \ell_{\theta, T}^{\chi, 0}(\mathbf{Y}) \xrightarrow{T \rightarrow +\infty} \ln W(\theta).$$

Therefore, when $T \rightarrow +\infty$, (22) implies

$$\ln(W(\mathbf{R}(\theta))/W(\theta)) \geq \phi(\mathbf{R}(\theta)) + \langle \bar{S}(\theta), \psi(\mathbf{R}(\theta)) \rangle - \phi(\theta) - \langle \bar{S}(\theta), \psi(\theta) \rangle. \quad (23)$$

By definition of $\bar{\theta}$ and \mathbf{R} (see A1(c) and (9)), the RHS is non negative. This concludes the proof of Proposition 4.2(i).

Proof of Proposition 4.2 (ii) We prove that $W \circ \mathbf{R}(\theta) - W(\theta) = 0$ if and only if $\theta \in \mathcal{L}$. Since $W \circ \mathbf{R} - W$ is continuous, this implies that $\inf_{\theta \in \mathcal{K}} W \circ \mathbf{R}(\theta) - W(\theta) > 0$ for all compact set $\mathcal{K} \subset \Theta \setminus \mathcal{L}$. Let $\theta \in \Theta$ be s.t. $W \circ \mathbf{R}(\theta) - W(\theta) = 0$. Then, the RHS in (23) is equal to zero. By definition of $\bar{\theta}$, $\mathbf{R}(\theta) = \theta$ and thus $\theta \in \mathcal{L}$. The converse implication is immediate from the definition of \mathcal{L} .

6.3 Proof of Theorem 4.4

The proof of Theorem 4.4 follows the same lines as the proof of [17, Theorem 3]. The key ingredient for this proof is the control of the L_p -mean error between the Block Online EM algorithm and the *limiting EM*. This is the crucial difference with [17]. The proof of this bound is derived in Proposition 6.5 and relies on preliminary lemmas; the detailed proof of Theorem 4.4 is given in [21, Section 3.1].

In the sequel, for all function Ξ on $\Theta \times \mathbb{Y}^{\mathbb{Z}}$ and all $\theta_\star \in \Theta$, we denote by $\mathbb{E}_\star[\Xi(\theta, \mathbf{Y})]_{\theta=\theta_\star}$ the function $\theta \mapsto \mathbb{E}_\star[\Xi(\theta, \mathbf{Y})]$ evaluated at $\theta = \theta_\star$. Finally, for any $L \geq 1$, $m \geq 1$ and any distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, define

$$\kappa_{L, m}^\chi(\theta, \mathbf{Y}) \stackrel{\text{def}}{=} \Phi_{\theta, L, L+m}^{\chi, L-m}(S, \mathbf{Y}) - \mathbb{E}_\star[\Phi_{v, 0, m}^{\chi, -m}(S, \mathbf{Y})]_{v=\theta}. \quad (24)$$

Lemma 6.2. *Assume A2, A3-(\bar{p}), A4(a) and A5 for some $\bar{p} > 2$. Let $p \in (2, \bar{p})$. There exists a constant C s.t. for any distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, any $m \geq 1$, $k, \ell \geq 0$ and any Θ -valued $\mathcal{F}_0^{\mathbf{Y}}$ -measurable r.v. θ ,*

$$\left\| \sum_{u=1}^k \kappa_{2um+\ell, m}^\chi(\theta, \mathbf{Y}) \right\|_{\star, p} \leq C \left[\sqrt{\frac{k}{m}} + k\beta^m \Delta p \right],$$

where $\Delta p \stackrel{\text{def}}{=} \frac{\bar{p}-p}{p\bar{p}}$ and β is given by A5.

Proof. For ease of notation χ is dropped from the notation $\kappa_{2um,m}^\chi$. By the Berbee Lemma (see [27, Chapter 5]), for any $m \geq 1$, there exists a Θ -valued r.v. $\boldsymbol{\theta}^*$ on $(\Omega, \mathcal{F}, \mathbb{P}_*)$ independent from $\mathcal{G}_m^\mathbf{Y}$ (see Eq.(10)) s.t.

$$\mathbb{P}_* \{ \boldsymbol{\theta} \neq \boldsymbol{\theta}^* \} = \sup_{B \in \mathcal{G}_m^\mathbf{Y}} |\mathbb{P}_*(B|\sigma(\boldsymbol{\theta})) - \mathbb{P}_*(B)|. \quad (25)$$

Set $L_u \stackrel{\text{def}}{=} 2um + \ell$. We write

$$\begin{aligned} \sum_{u=1}^k \kappa_{L_u,m}(\boldsymbol{\theta}, \mathbf{Y}) &= \sum_{u=1}^k \left\{ \Phi_{\boldsymbol{\theta}, L_u, L_u+m}^{\chi, L_u-m}(S, \mathbf{Y}) - \Phi_{\boldsymbol{\theta}^*, L_u, L_u+m}^{\chi, L_u-m}(S, \mathbf{Y}) \right\} \\ &+ \sum_{u=1}^k \kappa_{L_u,m}(\boldsymbol{\theta}^*, \mathbf{Y}) + k \left\{ \mathbb{E}_* [\Phi_{v,0,m}^{\chi, -m}(S, \mathbf{Y})]_{v=\boldsymbol{\theta}^*} - \mathbb{E}_* [\Phi_{v,0,m}^{\chi, -m}(S, \mathbf{Y})]_{v=\boldsymbol{\theta}} \right\}. \end{aligned} \quad (26)$$

By the Holder's inequality with $a \stackrel{\text{def}}{=} \bar{p}/p$ and $b^{-1} \stackrel{\text{def}}{=} 1 - a^{-1}$,

$$\begin{aligned} &\left\| \Phi_{\boldsymbol{\theta}, L, L+m}^{\chi, L-m}(S, \mathbf{Y}) - \Phi_{\boldsymbol{\theta}^*, L, L+m}^{\chi, L-m}(S, \mathbf{Y}) \right\|_{*,p} \\ &\leq \left\| \Phi_{\boldsymbol{\theta}, L, L+m}^{\chi, L-m}(S, \vartheta^T \mathbf{Y}) - \Phi_{\boldsymbol{\theta}^*, L, L+m}^{\chi, L-m}(S, \mathbf{Y}) \right\|_{*,\bar{p}} \mathbb{P}_* \{ \boldsymbol{\theta} \neq \boldsymbol{\theta}^* \}^{\Delta p}. \end{aligned}$$

By A4(a), A3(\bar{p}), A5, (1), (25) and (11), there exists a constant C_1 s.t. for any $m, L \geq 1$, any distribution χ and any Θ -valued $\mathcal{F}_0^\mathbf{Y}$ -measurable r.v. $\boldsymbol{\theta}$,

$$\left\| \Phi_{\boldsymbol{\theta}, L, L+m}^{\chi, L-m}(S, \mathbf{Y}) - \Phi_{\boldsymbol{\theta}^*, L, L+m}^{\chi, L-m}(S, \mathbf{Y}) \right\|_{*,\bar{p}} \leq C_1 \beta^{m\Delta p}.$$

Similarly, there exists a constant C_2 s.t. for any $m \geq 1$, any distribution χ and any Θ -valued $\mathcal{F}_0^\mathbf{Y}$ -measurable r.v. $\boldsymbol{\theta}$,

$$\left\| \mathbb{E}_* [\Phi_{v,0,m}^{\chi, -m}(S, \mathbf{Y})]_{v=\boldsymbol{\theta}^*} - \mathbb{E}_* [\Phi_{v,0,m}^{\chi, -m}(S, \mathbf{Y})]_{v=\boldsymbol{\theta}} \right\|_{*,p} \leq C_2 \beta^{m\Delta p}.$$

Let us consider the second term in (26). For any $u \geq 1$ and any $v \in \Theta$, the r.v. $\kappa_{L_u,m}(v, \mathbf{Y})$ is a measurable function of \mathbf{Y}_i for all $L_u - m + 1 \leq i \leq L_u + m$. Since $L_u \geq 2um$, for any $v \in \Theta$, $\sum_{u=1}^k \kappa_{L_u,m}(v, \mathbf{Y})$ is $\mathcal{G}_m^\mathbf{Y}$ -measurable. $\boldsymbol{\theta}^*$ is independent from $\mathcal{G}_m^\mathbf{Y}$ so that:

$$\left\| \sum_{u=1}^k \kappa_{L_u,m}(\boldsymbol{\theta}^*, \mathbf{Y}) \right\|_{*,p} = \mathbb{E}_* \left[\mathbb{E}_* \left[\left| \sum_{u=1}^k \kappa_{L_u,m}(v, \mathbf{Y}) \right|^p \right]_{v=\boldsymbol{\theta}^*} \right]^{1/p}.$$

Define the strong mixing coefficient (see [7])

$$\alpha^\mathbf{Y}(r) \stackrel{\text{def}}{=} \sup_{u \in \mathbb{Z}} \sup_{(A,B) \in \mathcal{F}_u^\mathbf{Y} \times \mathcal{G}_{u+r}^\mathbf{Y}} |\mathbb{P}_*(A \cap B) - \mathbb{P}_*(A)\mathbb{P}_*(B)|, r \geq 0.$$

Then, [7, Theorem 14.1, p.210] implies that for any $m \geq 1$, the strong mixing coefficients of the sequence $\kappa_{(\mathbf{m})} \stackrel{\text{def}}{=} \{\kappa_{L_u, m}(v, \mathbf{Y})\}_{u \geq 1}$ satisfies $\alpha^{\kappa_{(\mathbf{m})}}(i) \leq \alpha^{\mathbf{Y}}(2(i-1)m+1)$. Furthermore, by [27, Theorem 2.5],

$$\left\| \sum_{u=1}^k \kappa_{L_u, m}(v, \mathbf{Y}) \right\|_{\star, p} \leq (2kp)^{1/2} \left(\int_0^1 [N_{(m)}(t) \wedge k]^{p/2} \mathcal{Q}_{v, m}^p(t) dt \right)^{1/p},$$

where $N_{(m)}(t) \stackrel{\text{def}}{=} \sum_{i \geq 1} \mathbf{1}_{\alpha^{\kappa_{(\mathbf{m})}}(i) > t}$ and $\mathcal{Q}_{v, m}$ denotes the inverse of the tail function $t \mapsto \mathbb{P}_{\star}(|\kappa_{L_u, m}(v, \mathbf{Y})| \geq t)$. The sequence \mathbf{Y} being stationary, this inverse function does not depend on u . By A5 and the inequality $\alpha^{\mathbf{Y}}(r) \leq \beta^{\mathbf{Y}}(r)$ (see e.g. [7, Chapter 13]), there exist $\beta \in [0, 1)$ and $C \in (0, 1)$ s.t. for any $u, m \geq 1$,

$$N_{(m)}(u) \leq \sum_{i \geq 1} \mathbf{1}_{\alpha^{\mathbf{Y}}(2(i-1)m+1) > u} \leq \sum_{i \geq 1} \mathbf{1}_{C\beta^{2(i-1)m} > u} \leq \left(\frac{\log u - \log C}{2m \log \beta} \right) \vee 0.$$

Let U be a uniform r.v. on $[0, 1]$. Observe that $C\beta^{2mb} < 1$. Then, by the Holder inequality applied with $a \stackrel{\text{def}}{=} \bar{p}/p$ and $b^{-1} \stackrel{\text{def}}{=} 1 - a^{-1}$,

$$\begin{aligned} & \left\| [N_{(m)}(U) \wedge k]^{1/2} \mathcal{Q}_{v, m}(U) \right\|_p \stackrel{\text{def}}{=} \left(\int_0^1 [N_{(m)}(u) \wedge k]^{p/2} \mathcal{Q}_{v, m}^p(u) du \right)^{1/p} \\ & \leq k^{1/2} \left\| \mathcal{Q}_{v, m}(U) \mathbf{1}_{U \leq C\beta^{2mk}} \right\|_p + \left[\frac{-1}{2m \log \beta} \right]^{1/2} \left\| \mathcal{Q}_{v, m}(U) \left(-\log \frac{U}{C} \right)^{1/2} \right\|_p \\ & \leq \left\{ (C\beta^{2mk})^{\Delta p} k^{1/2} + \left[\frac{-1}{2m \log \beta} \right]^{1/2} \left\| \left(-\log \frac{U}{C} \right)^{1/2} \right\|_{pb} \right\} \left\| \mathcal{Q}_{v, m}(U) \right\|_{\bar{p}}. \end{aligned}$$

Since U is uniform on $[0, 1]$, $\mathcal{Q}_{v, m}(U)$ and $|\kappa_{L_u, m}(v, \mathbf{Y})|$ have the same distribution, see [27]. Then, by [21, Lemma 4.5] and A3-(\bar{p}), there exists a constant C s.t. for any $v \in \Theta$, any $m \geq 1$,

$$\sup_{v \in \Theta} \left\| \mathcal{Q}_{v, m}(U) \right\|_{\bar{p}} \leq C \left\| \sup_{x, x' \in \mathbb{X}^2} |S(x, x', \mathbf{Y}_0)| \right\|_{\star, \bar{p}},$$

which concludes the proof. \square

Lemma 6.3. *Assume A2, A3-(\bar{p}), A4(a) and A5 for some $\bar{p} > 2$. Let $p \in (2, \bar{p})$. There exists a constant C s.t. for any $n \geq 1$, any $1 \leq m_n \leq \tau_{n+1}$ and any distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$,*

$$\left\| \frac{1}{\tau_{n+1}} \sum_{t=2m_n}^{2v_n m_n} \kappa_{t, m_n}^{\chi}(\theta_n, \vartheta^{T_n} \mathbf{Y}) \right\|_{\star, p} \leq C \left[\frac{1}{\sqrt{\tau_{n+1}}} + \beta^{m_n \Delta p} \right],$$

where $\kappa_{L, m}^{\chi}$ and β are defined by (24) and A5, $v_n \stackrel{\text{def}}{=} \left\lfloor \frac{\tau_{n+1}}{2m_n} \right\rfloor$ and $\Delta p \stackrel{\text{def}}{=} \frac{\bar{p}-p}{p\bar{p}}$.

Proof. We write,

$$\left\| \sum_{t=2m_n}^{2v_n m_n} \kappa_{t,m_n}^\chi(\theta_n, \vartheta^{T_n} \mathbf{Y}) \right\|_{\star,p} \leq \sum_{\ell=0}^{2m_n-1} \left\| \sum_{u=1}^{v_n-1} \kappa_{2um_n+\ell,m_n}^\chi(\theta_n, \vartheta^{T_n} \mathbf{Y}) \right\|_{\star,p}.$$

Observe that by definition $\theta_n \in \mathcal{F}_{T_n}^{\mathbf{Y}}$. Then, by Lemma 6.2, there exists a constant C s.t. for any $m_n \geq 1$ and any $\ell \geq 0$,

$$\left\| \sum_{u=1}^{v_n-1} \kappa_{2um_n+\ell,m_n}^\chi(\theta_n, \vartheta^{T_n} \mathbf{Y}) \right\|_{\star,p} \leq C \left[\sqrt{\frac{v_n}{m_n}} + v_n \beta^{m_n \Delta p} \right].$$

The proof is concluded upon noting that $\tau_{n+1} \geq 2m_n v_n$. \square

Lemma 6.4. Assume A2, A3-(\bar{p}) and A4(a) for some $\bar{p} > 2$. For any $p \in (2, \bar{p}]$, there exists a constant C s.t. for any $n \geq 1$, any $1 \leq m_n \leq q_n \leq \tau_{n+1}$ and any distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$,

$$\left\| \bar{S}_{\tau_{n+1}}^{\chi, T_n}(\theta_n, \mathbf{Y}) - \bar{S}(\theta_n) - \tilde{\rho}_n \right\|_{\star,p} \leq C \left[\rho^{m_n \wedge (\tau_{n+1} - m_n)} + \frac{m_n}{\tau_{n+1}} + \frac{\tau_{n+1} - q_n}{\tau_{n+1}} \right],$$

where $\tilde{\rho}_n \stackrel{\text{def}}{=} \frac{1}{\tau_{n+1}} \sum_{t=2m_n}^{q_n} \kappa_{t,m_n}^\chi(\theta_n, \vartheta^{T_n} \mathbf{Y})$ and $\kappa_{L,m}^\chi$ is defined by (24).

Proof. By (1) and (3), $\bar{S}_{\tau_{n+1}}^{\chi, T_n}(\theta_n, \mathbf{Y}) - \bar{S}(\theta_n) - \tilde{\rho}_n = \sum_{i=1}^4 g_{i,n}$ where

$$\begin{aligned} g_{1,n} &\stackrel{\text{def}}{=} \frac{1}{\tau_{n+1}} \sum_{t=1}^{\tau_{n+1}} \Phi_{\theta_n, t, \tau_{n+1}}^{\chi, 0}(S, \vartheta^{T_n} \mathbf{Y}) - \frac{1}{\tau_{n+1}} \sum_{t=1}^{\tau_{n+1}} \Phi_{\theta_n, t, t+m_n}^{\chi, t-m_n}(S, \vartheta^{T_n} \mathbf{Y}), \\ g_{2,n} &\stackrel{\text{def}}{=} \frac{1}{\tau_{n+1}} \sum_{t=1}^{2m_n-1} \left(\Phi_{\theta_n, t, t+m_n}^{\chi, t-m_n}(S, \vartheta^{T_n} \mathbf{Y}) - \mathbb{E}_\star \left[\Phi_{\theta, 0, m_n}^{\chi, -m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} \right), \\ g_{3,n} &\stackrel{\text{def}}{=} \frac{1}{\tau_{n+1}} \sum_{t=q_n+1}^{\tau_{n+1}} \left(\Phi_{\theta_n, t, t+m_n}^{\chi, t-m_n}(S, \vartheta^{T_n} \mathbf{Y}) - \mathbb{E}_\star \left[\Phi_{\theta, 0, m_n}^{\chi, -m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} \right), \\ g_{4,n} &\stackrel{\text{def}}{=} \mathbb{E}_\star \left[\Phi_{\theta, 0, m_n}^{\chi, -m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} - \bar{S}(\theta_n). \end{aligned}$$

In the case $\tau_{n+1} > 2m_n$, it holds

$$\begin{aligned} \tau_{n+1} |g_{1,n}| &\leq \sum_{t=\tau_{n+1}-m_n+1}^{\tau_{n+1}} (\rho^{m_n-1} + \rho^{\tau_{n+1}-t}) \text{osc}\{S(\cdot, \cdot, Y_{t+T_n})\} \\ &+ \sum_{t=1}^{m_n} (\rho^{m_n} + \rho^{t-1}) \text{osc}\{S(\cdot, \cdot, Y_{t+T_n})\} + 2\rho^{m_n-1} \sum_{t=m_n+1}^{\tau_{n+1}-m_n} \text{osc}\{S(\cdot, \cdot, Y_{t+T_n})\}, \end{aligned}$$

where we used [21, Proposition 4.3(i) and Remark 4.4] in the last inequality. By A3-(\bar{p}) and A4(a), there exists C s.t. $\|g_{1,n}\|_{\star,p} \leq C(\rho^{m_n} + \tau_{n+1}^{-1})$. In the

case $\tau_{n+1} \leq 2m_n$, it can be proved along the same lines that $\|g_{1,n}\|_{\star,p} \leq C (\rho^{\tau_{n+1}-m_n} + \tau_{n+1}^{-1})$. For $g_{2,n}$ and $g_{3,n}$, we use the bounds

$$\begin{aligned} & \left| \Phi_{\theta_n, t, t+m_n}^{\chi, t-m_n}(S, \vartheta^{T_n} \mathbf{Y}) - \mathbb{E}_{\star} \left[\Phi_{\theta, 0, m_n}^{\chi, -m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} \right| \\ & \leq \sup_{(x, x') \in \mathbb{X}^2} |S(x, x', Y_{T_n+t})| + \mathbb{E}_{\star} \left[\sup_{(x, x') \in \mathbb{X}^2} |S(x, x', Y_0)| \right]. \end{aligned}$$

Then, by A4(a),

$$\begin{aligned} & \left\| \Phi_{\theta_n, t, t+m_n}^{\chi, t-m_n}(S, \vartheta^{T_n} \mathbf{Y}) - \mathbb{E}_{\star} \left[\Phi_{\theta, 0, m_n}^{\chi, -m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} \right\|_{\star,p} \\ & \leq 2 \left\| \sup_{(x, x') \in \mathbb{X}^2} |S(x, x', Y_0)| \right\|_{\star,p}, \end{aligned}$$

and the RHS is finite under A3-(\bar{p}). Finally,

$$|g_{4,n}| \leq 2\rho^{m_n-1} \mathbb{E}_{\star} [\text{osc}\{S(\cdot, \cdot, Y_0)\}] ,$$

where we used Theorem 4.1. This concludes the proof. \square

Proposition 6.5. *Assume A2, A3-(\bar{p}), A4(a) and A5 for some $\bar{p} > 2$. Let $p \in (2, \bar{p})$. There exists a constant C s.t. for any $n \geq 1$ and any distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$,*

$$\left\| \bar{S}_{\tau_{n+1}}^{\chi, T_n}(\theta_n, \mathbf{Y}) - \bar{S}(\theta_n) \right\|_{\star,p} \leq \frac{C}{\sqrt{\tau_{n+1}}} .$$

Proof. Let m_n, v_n be positive integers s.t. $1 \leq m_n \leq \tau_{n+1}$ and $\tau_{n+1} = 2v_n m_n + r_n$, where $0 \leq r_n < 2m_n$. Set $\Delta p \stackrel{\text{def}}{=} 1/p - 1/\bar{p}$. By the Minkowski inequality combined with Lemmas 6.3, 6.4 applied with $q_n \stackrel{\text{def}}{=} 2v_n m_n$, there exists a constant C s.t.

$$\left\| \bar{S}_{\tau_{n+1}}^{\chi, T_n}(\theta_n, \mathbf{Y}) - \bar{S}(\theta_n) \right\|_{\star,p} \leq C \left[\rho^{m_n} + \frac{m_n}{\tau_{n+1}} + \beta^{m_n \Delta p} + \frac{1}{\sqrt{\tau_{n+1}}} \right] .$$

The proof is concluded by choosing $m_n = \lfloor -\log \tau_{n+1} / (\log \rho \vee \Delta p \log \beta) \rfloor$. \square

6.4 Proof of Section 5

6.4.1 Proof of Theorem 5.1, Eq. (16)

Define $s_{\star} \stackrel{\text{def}}{=} \bar{S}(\theta_{\star})$ and

$$S_0 \stackrel{\text{def}}{=} \bar{S}_{\tau_1}^{\chi, 0}(\theta_0, \mathbf{Y}) \quad \text{and} \quad S_n \stackrel{\text{def}}{=} \bar{S}_{\tau_{n+1}}^{\chi, T_n}(\theta_n, \mathbf{Y}), \quad \forall n \geq 0. \quad (27)$$

We have:

$$\sqrt{\tau_n}(\theta_n - \theta_\star) = \sqrt{\tau_n}(\bar{\theta}(S_n) - \bar{\theta}(s_\star)) \mathbf{1}_{\bar{\theta}(S_n) \in \Theta_{p_n}} + \sqrt{\tau_n}(\theta_0 - \bar{\theta}(s_\star)) \mathbf{1}_{\bar{\theta}(S_n) \notin \Theta_{p_n}},$$

where p_n is defined by (6) and $\{\Theta_n\}_{n \geq 0}$ is defined by (5). By Theorem 4.4(a), the number of truncations is a.s. finite so that the second term is $o_{a.s.}(1)$. We write, for the first term,

$$\bar{\theta}(S_n) - \bar{\theta}(s_\star) = \Upsilon(S_n - s_\star) + \bar{\theta}(S_n) - \bar{\theta}(s_\star) - \Upsilon(S_n - s_\star), \quad (28)$$

where $\Upsilon \stackrel{\text{def}}{=} \nabla \bar{\theta}(s_\star)$. We now derive the rate of convergence of the quantity $S_n - s_\star$. Set $G(s) \stackrel{\text{def}}{=} \bar{S} \circ \bar{\theta}(s)$. Note that under A7(b), $\text{sp}(\Gamma) \leq \gamma$, where $\Gamma \stackrel{\text{def}}{=} \nabla G(s_\star)$, since $\Gamma = \nabla \bar{S}(\theta_\star) \cdot \nabla \bar{\theta}(s_\star)$. Since $G(s_\star) = s_\star$, we write

$$S_n - s_\star = \Gamma(S_{n-1} - s_\star) + S_n - G(S_{n-1}) + G(S_{n-1}) - G(s_\star) - \Gamma(S_{n-1} - s_\star).$$

Define $\{\mu_n\}_{n \geq 0}$ and $\{\rho_n\}_{n \geq 0}$ s.t. $\mu_0 = 0$, $\rho_0 = S_0 - s_\star$ and

$$\mu_n \stackrel{\text{def}}{=} \Gamma \mu_{n-1} + e_n, \quad \rho_n \stackrel{\text{def}}{=} S_n - s_\star - \mu_n, \quad n \geq 1, \quad (29)$$

where,

$$e_n \stackrel{\text{def}}{=} S_n - \bar{S}(\theta_n), \quad n \geq 1. \quad (30)$$

Proposition 6.6. *Assume A2, A3-(\bar{p}_2), A4-5, A6-(\bar{p}_1), A7 and A8(a) for some $2 < \bar{p}_1 < \bar{p}_2$. Then for any $p \in (2, \bar{p}_2)$, $\sqrt{\tau_n} \mu_n = \mathcal{O}_{L_p}(1)$ and $\tau_k \rho_k \mathbf{1}_{\lim_n S_n = s_\star} = \mathcal{O}_{L_{p/2}}(1) \mathcal{O}_{a.s.}(1)$.*

The proof of Proposition 6.6 is on the same lines as the proof of [17, Theorem 6]. The main ingredient is the control of $\|\mu_n\|_{\star, p}$ which is a consequence of [25, Result 178, p. 39] and Proposition 6.5. The detailed proof is thus omitted and postponed to the supplementary material [21, Section 3.2].

By Proposition 6.6, the first term in (28) gives

$$\sqrt{\tau_n} \Upsilon(S_n - s_\star) \mathbf{1}_{\lim_n \theta_n = \theta_\star} \mathbf{1}_{\bar{\theta}(S_n) \in \Theta_{p_n}} = \mathcal{O}_{L_p}(1) + \frac{1}{\sqrt{\tau_n}} \mathcal{O}_{L_{p/2}}(1) \mathcal{O}_{a.s.}(1).$$

A Taylor expansion with integral remainder term gives the rate of convergence of the second term.

6.4.2 Proof of Theorem 5.1, Eq.(17)

We preface the proof by the following lemma.

Lemma 6.7. *Assume A2, A3-(\bar{p}_2), A4-5, A7, A8(b) for some $\bar{p}_2 > 2$. For any $p \in (2, \bar{p}_2)$,*

$$\limsup_{n \rightarrow +\infty} \frac{1}{\sqrt{T_{n+1}}} \left\| \sum_{k=1}^n \tau_{k+1} e_k \right\|_{\star, p} < \infty,$$

where e_n is given by (30).

Proof. Let $p \in (2, \bar{p}_2)$. In the sequel, C is a constant independent on n and whose value may change upon each appearance. Let $1 \leq m_n \leq \tau_{n+1}$ and set $v_n \stackrel{\text{def}}{=} \left\lfloor \frac{\tau_{n+1}}{2m_n} \right\rfloor$. By Lemma 6.4 applied with $q_k \stackrel{\text{def}}{=} 2v_k m_k$, we have,

$$\left\| \sum_{k=1}^n \tau_{k+1} e_k \right\|_{\star, p} \leq C \left(\sum_{k=1}^n \{ \tau_{k+1} \rho^{m_k \wedge (\tau_{k+1} - m_k)} + m_k \} + \left\| \sum_{k=1}^n \{ \delta_k + \zeta_k \} \right\|_{\star, p} \right),$$

where δ_k and ζ_k are defined by

$$\begin{aligned} \delta_k &\stackrel{\text{def}}{=} \sum_{t=2m_k}^{2v_k m_k} \{ F_{t,k}(\theta_k, \mathbf{Y}) - \mathbb{E}_\star [F_{t,k}(\theta_k, \mathbf{Y}) | \mathcal{F}_{T_k}^{\mathbf{Y}}] \}, \\ \zeta_k &\stackrel{\text{def}}{=} \sum_{t=2m_k}^{2v_k m_k} \left\{ \mathbb{E}_\star [F_{t,k}(\theta_k, \mathbf{Y}) | \mathcal{F}_{T_k}^{\mathbf{Y}}] - \mathbb{E}_\star [\Phi_{\theta,0,m_k}^{\chi,-m_k}(S, \mathbf{Y})]_{\theta=\theta_k} \right\}, \end{aligned}$$

where $F_{t,k}(\theta_k, \mathbf{Y}) \stackrel{\text{def}}{=} \Phi_{\theta_k,t,t+m_k}^{\chi,-m_k}(S, \vartheta^{T_k} \mathbf{Y})$ and $\mathcal{F}_{T_k}^{\mathbf{Y}}$ is given by (10). We will prove below that there exists C s.t.

$$\|\zeta_k\|_{\star, p} \leq C \beta^{m_k/pb} \tau_{k+1}, \quad \forall k \geq 1 \quad (31)$$

$$\left\| \sum_{k=1}^n \delta_k \right\|_{\star, p} \leq C \sqrt{T_{n+1}} + C \sum_{k=1}^n \tau_{k+1} \beta^{m_k/pb}, \quad \forall n \geq 1 \quad (32)$$

so that the proof is concluded by choosing $m_k = \lfloor \eta \log \tau_{k+1} \rfloor$, $\eta \stackrel{\text{def}}{=} (-1/\log \rho) \vee (-pb/\log \beta)$.

We turn to the proof of (31). By the Berbee Lemma (see [27, Chapter 5]) and A5, there exist $C \in [0, 1)$ and $\beta \in (0, 1)$ s.t. for all $k \geq 1$, there exists a random variable $Y_{T_k+m_k:T_{k+1}+m_k}^{\star, (k)}$ on $(\Omega, \mathcal{F}, \mathbb{P}_\star)$ independent from $\mathcal{F}_{T_k}^{\mathbf{Y}}$ with the same distribution as $Y_{T_k+m_k:T_{k+1}+m_k}$ and

$$\mathbb{P}_\star \left\{ Y_{T_k+m_k:T_{k+1}+m_k}^{\star, (k)} \neq Y_{T_k+m_k:T_{k+1}+m_k} \right\} \leq C \beta^{m_k}. \quad (33)$$

Upon noting that $\mathbb{E}_\star [F_{t,k}(\theta_k, \mathbf{Y}^{\star, (k)}) | \mathcal{F}_{T_k}^{\mathbf{Y}}] = \mathbb{E}_\star [F_{t,k}(\theta, \mathbf{Y})]_{\theta=\theta_k}$, we have

$$\zeta_k = \sum_{t=2m_k}^{2v_k m_k} \left\{ \mathbb{E}_\star [F_{t,k}(\theta_k, \mathbf{Y}) | \mathcal{F}_{T_k}^{\mathbf{Y}}] - \mathbb{E}_\star [F_{t,k}(\theta_k, \mathbf{Y}^{\star, (k)}) | \mathcal{F}_{T_k}^{\mathbf{Y}}] \right\}. \quad (34)$$

Therefore, by setting $\mathcal{A}_k \stackrel{\text{def}}{=} \{ Y_{T_k+m_k:T_{k+1}+m_k}^{\star, (k)} \neq Y_{T_k+m_k:T_{k+1}+m_k} \}$,

$$|\zeta_k| \leq \sum_{t=2m_k}^{2v_k m_k} \mathbb{E}_\star \left[\sup_{\theta \in \Theta} \left| F_{t,k}(\theta, \mathbf{Y}) - F_{t,k}(\theta, \mathbf{Y}^{\star, (k)}) \right| \mathbf{1}_{\mathcal{A}_k} \middle| \mathcal{F}_{T_k}^{\mathbf{Y}} \right].$$

Minkowski and Holder (with $a \stackrel{\text{def}}{=} \bar{p}_2/p$ and $b^{-1} \stackrel{\text{def}}{=} 1 - a^{-1}$) inequalities, combined with (33), A4(a), [21, Lemma 4.5] and A3-(\bar{p}_2) yield (31).

We now prove (32). Upon noting that δ_k is $\mathcal{F}_{T_{k+1}}^{\mathbf{Y}}$ -measurable and δ_k is a martingale increment, the Rosenthal inequality (see [18, Theorem 2.12, p.23]) states that $\|\sum_{k=1}^n \delta_k\|_{\star,p} \leq C \left(\sum_{k=1}^n I_k^{(1)} \right)^{1/p} + CI_n^{(2)}$ where

$$I_k^{(1)} \stackrel{\text{def}}{=} \mathbb{E}_{\star} [|\delta_k|^p] \quad \text{and} \quad I_n^{(2)} \stackrel{\text{def}}{=} \left\| \left(\sum_{k=1}^n \mathbb{E}_{\star} [|\delta_k|^2 | \mathcal{F}_{T_k}^{\mathbf{Y}}] \right)^{1/2} \right\|_{\star,p}.$$

Using again $\mathbb{E}_{\star} [F_{t,k}(\theta_k, \mathbf{Y}^{*,(k)}) | \mathcal{F}_{T_k}^{\mathbf{Y}}] = \mathbb{E}_{\star} [F_{t,k}(\theta, \mathbf{Y})]_{\theta=\theta_k}$ and (34)

$$I_k^{(1)} \leq C \left\| \sum_{t=2m_k}^{2v_k m_k} \left\{ F_{t,k}(\theta_k, \mathbf{Y}) - \mathbb{E}_{\star} [F_{t,k}(\theta, \mathbf{Y})]_{\theta=\theta_k} \right\} \right\|_{\star,p}^p + C \|\zeta_k\|_{\star,p}^p.$$

By Lemma 6.3 and (31), there exists C s.t. for any $k \geq 1$

$$I_k^{(1)} \leq C \left(\tau_{k+1}^{p/2} + \tau_{k+1}^p \beta^{m_k/b} \right), \quad (35)$$

and since $2/p < 1$, convex inequalities yield $\left(\sum_{k=1}^n I_k^{(1)} \right)^{1/p} \leq C \sqrt{T_{n+1}} + C \sum_{k=1}^n \tau_{k+1} \beta^{m_k/pb}$. By the Minkowski and Jensen inequalities, it holds $I_n^{(2)} \leq \left(\sum_{k=1}^n \{I_k^{(1)}\}^{2/p} \right)^{1/2}$. Hence, by (35), $I_n^{(2)} \leq C \sqrt{T_{n+1}} + C \sum_{k=1}^n \tau_{k+1} \beta^{m_k/pb}$. This concludes the proof of (32). \square

We write $\Sigma_n - s_{\star} = \bar{\mu}_n + \bar{\rho}_n$ with

$$\bar{\mu}_n \stackrel{\text{def}}{=} \frac{1}{T_n} \sum_{k=1}^n \tau_k \mu_{k-1} \quad \text{and} \quad \bar{\rho}_n \stackrel{\text{def}}{=} \frac{1}{T_n} \sum_{k=1}^n \tau_k \rho_{k-1}. \quad (36)$$

Proposition 6.8. *Assume A2, A3-(\bar{p}_2), A4-5, A6-(\bar{p}_1), A7 and A8 for some $2 < \bar{p}_1 < \bar{p}_2$. For any $p \in (2, \bar{p}_2)$,*

$$\sqrt{T_n} \bar{\mu}_n = \mathcal{O}_{L_p}(1), \quad \frac{T_n}{n} \bar{\rho}_n \mathbf{1}_{\lim_n S_n = s_{\star}} = \mathcal{O}_{L_{p/2}}(1) \mathcal{O}_{\text{a.s.}}(1).$$

Proof. Set $A \stackrel{\text{def}}{=} (I - q\Gamma)$. Under A7, A^{-1} exists. By (29) and (36),

$$A \sqrt{T_n} \bar{\mu}_n = -\frac{\tau_{n+1} \mu_n}{\sqrt{T_n}} + \frac{1}{\sqrt{T_n}} \sum_{k=1}^n \tau_{k+1} e_k + \frac{1}{\sqrt{T_n}} \sum_{k=1}^n \tau_k \left(\frac{\tau_{k+1}}{\tau_k} - q \right) \Gamma \mu_{k-1}.$$

The result now follows from Proposition 6.6, Lemma 6.7 and A8. The proof of the second assertion follows from (36) and Proposition 6.6. \square

Upon noting that $\theta_\star = \bar{\theta}(s_\star)$, we may write, for the averaged sequence,

$$\tilde{\theta}_n - \theta_\star = \Upsilon(\Sigma_n - s_\star) + \bar{\theta}(\Sigma_n) - \bar{\theta}(s_\star) - \Upsilon(\Sigma_n - s_\star) .$$

The first term in this decomposition gives

$$\sqrt{T_n} \Upsilon(\Sigma_n - s_\star) \mathbf{1}_{\lim_n \theta_n = \theta_\star} = \mathcal{O}_{L_p}(1) + \frac{n}{\sqrt{T_n}} \mathcal{O}_{L_{p/2}}(1) \mathcal{O}_{a.s.}(1) .$$

By A7(b), as for the non averaged sequence, a Taylor expansion with integral remainder term gives the result for the second term.

7 Acknowledgments

The authors are grateful to Eric Moulines and Olivier Cappé for their fruitful remarks.

References

- [1] P. Billingsley. *Probability and Measure*. Wiley, New York, 3rd edition, 1995.
- [2] O. Cappé. Online sequential Monte Carlo EM algorithm. In *IEEE Workshop on Statistical Signal Processing (SSP)*, 2009.
- [3] O. Cappé. Online EM algorithm for Hidden Markov Models. *J. Comput. Graph. Statist.*, 20(3):728–749, 2011.
- [4] O. Cappé and E. Moulines. Online Expectation Maximization algorithm for latent data models. *J. Roy. Statist. Soc. B*, 71(3):593–613, 2009.
- [5] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.
- [6] H. Chen, A. Gao, and L. Guo. Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stoch. Proc. Appl.*, 27:217–231, 1988.
- [7] J. Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, 1994.
- [8] M. Del Moral, A. Doucet, and S.S Singh. Forward smoothing using sequential Monte Carlo. arXiv:1012.5390v1, Dec 2010.
- [9] P. Del Moral and A. Guionnet. Large deviations for interacting particle systems: applications to non-linear filtering. *Stoch. Proc. Appl.*, 78:69–95, 1998.
- [10] P. Del Moral, M. Ledoux, and L. Miclo. On contraction properties of Markov kernels. *Probab. Theory Related Fields*, 126(3):395–420, 2003.

- [11] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1), 1999.
- [12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39(1):1–38 (with discussion), 1977.
- [13] R. Douc, G. Fort, E. Moulines, and P. Priouret. Forgetting the initial distribution for hidden Markov models. *Stochastic Processes and their Applications*, 119(4):1235–1256, 2009.
- [14] R. Douc, E. Moulines, and T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.*, 32(5):2254–2304, 2004.
- [15] J. Durbin and S. J. Koopman. Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *J. Roy. Statist. Soc. B*, 62:3–29, 2000.
- [16] Y. Ephraim and N. Merhav. Hidden Markov Processes. *IEEE Trans. on information theory*, 18(6):1518–1570, 2002.
- [17] G. Fort and E. Moulines. Convergence of the Monte Carlo Expectation Maximization for curved exponential families. *Ann. Statist.*, 31(4):1220–1259, 2003.
- [18] P. Hall and C. C. Heyde. *Martingale Limit Theory and its Application*. Academic Press, New York, London, 1980.
- [19] H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer, 1997.
- [20] S. Le Corff and G. Fort. Convergence of a particle-based approximation of the block online Expectation Maximization algorithm. Technical report, arXiv:1111.1307, 2011.
- [21] S. Le Corff and G. Fort. Supplementary to "Online Expectation Maximization based algorithms for inference in Hidden Markov Models". Technical report, arXiv:1108.4130, 2011.
- [22] S. Le Corff, G. Fort, and E. Moulines. Online EM algorithm to solve the SLAM problem. In *IEEE Workshop on Statistical Signal Processing (SSP)*, 2011.
- [23] F. Le Gland and L. Mevel. Recursive estimation in HMMs. In *Proc. IEEE Conf. Decis. Control*, pages 3468–3473, 1997.
- [24] G. Mongillo and S. Denève. Online learning with hidden Markov models. *Neural Computation*, 20(7):1706–1716, 2008.

- [25] G. Pólya and G. Szegő. *Problems and Theorems in Analysis. Vol. II.* Springer, 1976.
- [26] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.
- [27] E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants.* Springer, 1990.
- [28] Vladislav B. Tadić. Analyticity, convergence, and convergence rate of recursive maximum-likelihood estimation in hidden Markov models. *IEEE Trans. Inf. Theor.*, 56:6406–6432, December 2010.
- [29] D. M. Titterton. Recursive parameter estimation using incomplete data. *J. Roy. Statist. Soc. B*, 46(2):257–267, 1984.
- [30] R. Van Handel. Uniform time average consistency of Monte Carlo particle filters. *Stoch. Proc. Appl.*, 119:3835–3861, 2009.
- [31] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models.* Springer, 1989.
- [32] C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Statist.*, 11:95–103, 1983.